

The Quality of Content in Open Online Collaboration Platforms

Approaches to NLP-supported Information Quality Management in Wikipedia

Vom Fachbereich Informatik
der Technischen Universität Darmstadt
genehmigte

Dissertation

zur Erlangung des akademischen Grades
Doktor-Ingenieur (Dr.-Ing.)

vorgelegt von

Oliver Ferschke, M.A.
geboren in Würzburg

Tag der Einreichung: 15. Mai 2014

Tag der Disputation: 15. Juli 2014

Referenten: Prof. Dr. Iryna Gurevych, Darmstadt
Prof. Dr. Hinrich Schütze, München
Assoc. Prof. Carolyn Penstein Rosé, PhD, Pittsburgh

Darmstadt 2014

D17

Please cite this document as

URN: urn:nbn:de:tuda-tuprints-40929

URL: <http://tuprints.ulb.tu-darmstadt.de/4092>

This document is provided by tuprints,

E-Publishing-Service of the TU Darmstadt

<http://tuprints.ulb.tu-darmstadt.de>

tuprints@ulb.tu-darmstadt.de



This work is published under the following Creative Commons license:

Attribution – Non Commercial – No Derivative Works 2.0 Germany

<http://creativecommons.org/licenses/by-nc-nd/2.0/de/deed.en>

Abstract

Over the past decade, the paradigm of the World Wide Web has shifted from static web pages towards participatory and collaborative content production. The main properties of this user generated content are a low publication threshold and little or no editorial control. While this has improved the variety and timeliness of the available information, it causes an even higher variance in quality than the already heterogeneous quality of traditional web content. Wikipedia is the prime example for a successful, large-scale, collaboratively-created resource that reflects the spirit of the open collaborative content creation paradigm. Even though recent studies have confirmed that the overall quality of Wikipedia is high, there is still a wide gap that must be bridged before Wikipedia reaches the state of a reliable, citable source.

A key prerequisite to reaching this goal is a quality management strategy that can cope both with the massive scale of Wikipedia and its open and almost anarchic nature. This includes an efficient communication platform for work coordination among the collaborators as well as techniques for monitoring quality problems across the encyclopedia. This dissertation shows how natural language processing approaches can be used to assist information quality management on a massive scale.

In the first part of this thesis, we establish the theoretical foundations for our work. We first introduce the relatively new concept of *open* online collaboration with a particular focus on collaborative writing and proceed with a detailed discussion of Wikipedia and its role as an encyclopedia, a community, an online collaboration platform, and a knowledge resource for language technology applications. We then proceed with the three main contributions of this thesis.

Even though there have been previous attempts to adapt existing information quality frameworks to Wikipedia, no quality model has yet incorporated writing quality as a central factor. Since Wikipedia is not only a repository of mere facts but rather consists of full text articles, the writing quality of these articles has to be taken into consideration when judging article quality. As the first main contribution of this thesis, we therefore define a comprehensive article quality model that aims to consolidate both the quality of writing

and the quality criteria defined in multiple Wikipedia guidelines and policies into a single model. The model comprises 23 dimensions segmented into the four layers of intrinsic quality, contextual quality, writing quality and organizational quality.

As a second main contribution, we present an approach for automatically identifying quality flaws in Wikipedia articles. Even though the general idea of quality detection has been introduced in previous work, we dissect the approach to find that the task is inherently prone to a topic bias which results in unrealistically high cross-validated evaluation results that do not reflect the classifier’s real performance on real world data.

We solve this problem with a novel data sampling approach based on the full article revision history that is able to avoid this bias. It furthermore allows us not only to identify flawed articles but also to find reliable counterexamples that do not exhibit the respective quality flaws. For automatically detecting quality flaws in unseen articles, we present *Flaw-Finder*, a modular system for supervised text classification. We evaluate the system on a novel corpus of Wikipedia articles with neutrality and style flaws. The results confirm the initial hypothesis that the reliable classifiers tend to exhibit a lower cross-validated performance than the biased ones but the scores more closely resemble their actual performance in the wild.

As a third main contribution, we present an approach for automatically segmenting and tagging the user contributions on article Talk pages to improve the work coordination among Wikipedians. These unstructured discussion pages are not easy to navigate and information is likely to get lost over time in the discussion archives. By automatically identifying the quality problems that have been discussed in the past and the solutions that have been proposed, we can help users to make informed decisions in the future.

Our contribution in this area is threefold: (i) We describe a novel algorithm for segmenting the unstructured dialog on Wikipedia Talk pages using their revision history. In contrast to related work, which mainly relies on the rudimentary markup, this new algorithm can reliably extract meta data, such as the identity of a user, and is moreover able to handle discontinuous turns. (ii) We introduce a novel scheme for annotating the turns in article discussions with dialog act labels for capturing the coordination efforts of article improvement. The labels reflect the types of criticism discussed in a turn, for example missing information or inappropriate language, as well as any actions proposed for solving the quality problems. (iii) Based on this scheme, we created two automatically segmented and manually annotated discussion corpora extracted from the Simple English Wikipedia (SEWD) and the English Wikipedia (EWD). We evaluate how well text classification approaches can learn to assign the dialog act labels from our scheme to unseen discussion pages and achieve a cross-validated performance of $F_1 = 0.82$ on the SEWD corpus while we obtain an average performance of $F_1 = 0.78$ on the larger and more complex EWD corpus.

Zusammenfassung

In den vergangenen zehn Jahren hat sich der Fokus des World Wide Web von primär statischen Webseiten hin zu kollaborativ erstellten Inhalten verlagert. Die wichtigsten Eigenschaften dieses neuen Paradigmas sind eine niedrige Veröffentlichungsschwelle und wenig oder gänzlich fehlende redaktionelle Kontrolle. Wenngleich dadurch die Vielfalt und Aktualität der verfügbaren Informationen verbessert wurde, fördert es zugleich auch die Heterogenität der Webinhalte hinsichtlich ihrer Qualität. Wikipedia ist das Paradebeispiel für eine große, erfolgreiche, kollaborativ erstellte Ressource, die den Geist freier Kollaboration widerspiegelt. Auch wenn jüngste Studien bestätigt haben, dass die Qualität von Wikipedia insgesamt hoch ist, ist es immer noch ein weiter Weg Wikipedia zu einer zuverlässigen und zitierbaren Quelle zu machen.

Eine wichtige Voraussetzung zur Erreichung dieses Ziels ist eine Qualitätsmanagementstrategie, die sowohl mit der Größe von Wikipedia und ihrer offenen, nahezu anarchischen Organisationsstruktur umgehen kann. Eine solche Strategie schließt eine effiziente Kommunikationsplattform für die Arbeitskoordination zwischen den Nutzern, sowie Techniken zur Überwachung von Qualitätsproblemen in der Enzyklopädie mit ein. Diese Dissertation zeigt auf, wie sprachtechnologische Methoden die bestehenden Ansätze zum Informationsqualitätsmanagement in Wikipedia effektiv unterstützen können. Im ersten Teil der Dissertation führen wir die theoretischen Grundlagen für unsere Arbeit ein. Wir erörtern zunächst das relativ neue Konzept der *freien* Online-Kollaboration unter besonderer Berücksichtigung kollaborativen Schreibens. Vervollständigt wird diese Einführung mit einer ausführlichen Diskussion der Wikipedia. Auf Basis dieser Grundlagen folgen die drei Hauptbeiträge der vorliegenden Arbeit.

Wenngleich es bereits Versuche gab, bestehende Frameworks zur Erfassung von Informationsqualität an die Bedürfnisse der Wikipedia anzupassen, hat bisher kein Modell die Text- und Schreibqualität als zentralen Faktor berücksichtigt. Da Wikipedia jedoch nicht nur eine Ansammlung von Fakten ist, sondern aus Volltextartikeln besteht, ist der Text- und Schreibqualität dieser Artikel eine zentrale Rolle bei den Qualitätsbetrachtungen zuzuschreiben. Als ersten zentralen Beitrag dieser Dissertation definieren wir daher ein um-

fassendes Artikelqualitätsmodell, welches sowohl die Text- und Schreibqualität als auch die spezifischen Qualitätskriterien der Wikipedia in einem einzigen Modell zusammenführt. Es umfasst insgesamt 23 Qualitätsdimensionen in den Kategorien *intrinsische Qualität*, *kontextbezogene Qualität*, *Text- und Schreibqualität* und *strukturelle Qualität*.

Im zweiten zentralen Beitrag dieser Arbeit, stellen wir einen Ansatz zur automatischen Erkennung von Qualitätsmängeln in Wikipedia-Artikeln vor. Auch wenn die Idee hierzu bereits in früheren Arbeiten beschrieben wurde, haben wir in unseren Experimenten herausgefunden, dass dieser Ansatz von Natur aus anfällig für ein Themenbias ist, welches zu unrealistisch hohen Werten in der Kreuzvalidierung von Klassifikationsmodellen führt. Die tatsächliche Leistung auf realen Daten liegt weit unter den Ergebnissen, die in früheren Arbeiten berichtet wurden. Wir lösen dieses Problem mit einem neuen Samplingverfahren basierend auf der Artikelrevisionsgeschichte. Dieser Ansatz vermag es nicht nur fehlerhafte Artikel zu identifizieren, sondern auch zuverlässige Gegenbeispiele zu finden, die nicht die entsprechenden Qualitätsmängel aufweisen. Zur automatischen Erkennung von Qualitätsmängeln haben wir FlawFinder entwickelt, ein modulares System für überwachte Textklassifikation. Wir evaluieren das System auf einem Korpus aus Wikipedia-Artikeln mit Qualitätsmängeln in den Bereichen Neutralität und Stilistik. Die gewonnenen Ergebnisse bestätigen unsere Ausgangshypothese, dass auf ausgeglichenen Daten trainierte Klassifikatoren zwar zu einer geringeren kreuzvalidierten Leistung neigen, jedoch die tatsächliche Leistung in realen Anwendungsszenarien realistischer widerspiegeln.

Als dritten zentralen Beitrag dieser Arbeit, stellen wir einen Ansatz für die automatische Segmentierung und Klassifikation von Nutzerbeiträgen in Artikeldiskussionsseiten vor. Es hat sich gezeigt, dass Nutzer der Wikipedia Probleme haben, sich auf diesen unstrukturierten Diskussionsseiten zurechtzufinden und archivierte Informationen mit der Zeit nur noch schwer auffindbar sind. Indem wir automatisch die Qualitätsprobleme und Lösungsvorschläge identifizieren, die in vergangenen Diskussionen erörtert wurden, können wir den Nutzern helfen, fundierte Entscheidungen in der Zukunft zu treffen. Der Beitrag unterteilt sich in folgende drei Teile: (i) Wir beschreiben einen neuen Algorithmus zur Segmentierung des unstrukturierten Dialogs auf Wikipedia-Diskussionsseiten mit Hilfe ihrer Revisionsgeschichte. (ii) Wir stellen ein neuartiges Annotationsschema für Beiträge in Artikeldiskussionen vor. Die darin definierten Dialogakte spiegeln wider, welche Kritik an einem Artikel geäußert wurde, wie zum Beispiel fehlende Informationen oder unangemessene Sprache, und welche Lösungen vorgeschlagen wurden. (iii) Basierend auf diesem Schema haben wir zwei automatisch segmentierte und manuell annotierte Korpora aus Artikeln der Simple English Wikipedia (SEWD) und der englischen Wikipedia (EWD) erstellt. Wir nutzen diese Korpora um Klassifikationsmodelle zu trainieren um die Dialogakte in unbekanntem Diskussionsseiten identifizieren zu können. In unserer Evaluation erreichen wir auf dem SEWD Korpus eine Leistung von $F_1 = 0.82$, während wir auf dem komplexeren EWD Korpus durchschnittlich $F_1 = 0.78$ beobachten konnten.

Acknowledgements

I would like to thank my advisor, Iryna Gurevych, for her guidance, enthusiasm and encouragement during my nearly five years at the UKP lab. I am grateful for her patience and support and for all the time she spared despite her constantly busy schedule. I am also indebted to my committee members, Carolyn Rosé and Hinrich Schütze, for reviewing my work and providing valuable feedback.

I wholeheartedly thank Torsten Zesch, a most gifted researcher and inspiring mentor, who did not only guide me with my first steps as a researcher but who was also a constant source of ideas, wisdom and moral support.

I furthermore thank Christian Meyer for the countless stimulating discussions, his helpful feedback to my work and for the many theater visits that managed to take my mind off work for a while.

Without the tireless dedication of Richard Eckart de Castilho, I would have surrendered more often than I would like to admit to the technical challenges of my work.

Special thanks also goes to my closest collaborator, Johannes Daxenberger, with whom I had the pleasure of co-authoring several publications and who was the most knowledgeable person with whom to discuss anything related to Wikipedia.

I thank Christian Meyer, Emily Jamison, Johannes Daxenberger, Lucie Flekova and Lisa Beinborn for reviewing my thesis and providing me with helpful feedback and many invaluable ideas. I furthermore thank all colleagues at UKP for the many joyful and enlightening hours at the lab and all student research assistants who contributed to my work.

Finally, I would like to express my deepest gratitude to my family and friends. Without their support, I would not have been able to successfully go through the roller coaster ride of graduate school. I especially thank my parents for always having an open ear for any problems and my nephew Maximilian for cheering me up whenever I visit.

This work has been supported by the Hessian research excellence program “Landes-Offensive zur Entwicklung Wissenschaftlich-Ökonomischer Exzellenz” (LOEWE) as part of the research center “Digital Humanities”.

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 1.1 | Main Contributions | 3 |
| 1.2 | Publication Record | 5 |
| 1.3 | Thesis Organization | 6 |
| 1.4 | Terminology and Conventions | 7 |
| 2 | Open Online Collaboration | 9 |
| 2.1 | Open Online Collaboration | 9 |
| 2.2 | Collaborative Writing | 12 |
| 2.3 | Collaborative Online Writing Systems | 15 |
| 2.4 | Chapter Summary | 17 |
| 3 | Wikipedia | 19 |
| 3.1 | Overview | 19 |
| 3.2 | Structure and Organization | 22 |
| 3.2.1 | Namespaces and Naming Conventions | 23 |
| 3.2.2 | Organizational Structures | 25 |
| 3.2.3 | Inner Article Structure | 26 |
| 3.2.4 | Template System | 28 |
| 3.3 | Community | 29 |
| 3.3.1 | User Groups and Roles | 29 |
| 3.3.2 | Soft Security | 30 |
| 3.3.3 | Systemic Bias | 30 |
| 3.3.4 | WikiProjects | 31 |
| 3.4 | Revision History | 31 |
| 3.5 | User Discussions | 33 |
| 3.6 | Processing Wikipedia | 37 |
| 3.6.1 | Data Sources | 37 |

| | | |
|----------|--|-----------|
| 3.6.2 | Data Access | 39 |
| 3.7 | Other Wikimedia Projects | 41 |
| 3.8 | Chapter Summary | 43 |
| 4 | Information Quality | 45 |
| 4.1 | Information Quality | 45 |
| 4.2 | Text and Writing Quality | 51 |
| 4.3 | Quality Management Mechanisms in Wikipedia | 53 |
| 4.4 | An Article Quality Model for Wikipedia | 58 |
| 4.4.1 | Intrinsic Article Quality | 60 |
| 4.4.2 | Contextual Article Quality | 61 |
| 4.4.3 | Article Writing Quality | 61 |
| 4.4.4 | Organizational Article Quality | 63 |
| 4.5 | Chapter Summary | 63 |
| 5 | Quality Flaw Detection in Wikipedia Articles | 65 |
| 5.1 | Motivation and Overview | 66 |
| 5.2 | Quality Flaws in Wikipedia | 67 |
| 5.2.1 | Properties of Quality Flaws in Wikipedia | 67 |
| 5.2.2 | Definition of the Quality Flaw Detection Task | 70 |
| 5.2.3 | Reliability of Cleanup Templates as Quality Flaw Markers | 71 |
| 5.2.4 | Coverage of the Article Quality Model by Cleanup Templates | 72 |
| 5.2.5 | Quality Flaw Markers in Non-English Wikipedias | 74 |
| 5.3 | Quality Flaw Corpora | 74 |
| 5.3.1 | The CLEF Corpus | 75 |
| 5.3.2 | Reliability of Training Instances | 76 |
| 5.3.3 | Topic Bias | 77 |
| 5.3.4 | The NSTYLE Corpus | 79 |
| 5.4 | A System for Quality Flaw Detection | 85 |
| 5.4.1 | System Architecture | 87 |
| 5.4.2 | Features | 90 |
| 5.5 | Experiments | 94 |
| 5.5.1 | Experiment Setup and Optimization | 94 |
| 5.5.2 | Evaluation and Error Analysis | 98 |
| 5.6 | Mining Flaw Corrections from the Revision History | 104 |
| 5.7 | Limitations in the Predictability of Quality Flaws | 106 |
| 5.8 | Chapter Summary | 108 |

| | | |
|----------|--|------------|
| 6 | Dialog Analysis of Wikipedia Talk Pages | 109 |
| 6.1 | Motivation and Overview | 110 |
| 6.2 | Linguistic Background | 111 |
| 6.3 | Related Work | 113 |
| 6.3.1 | Work Coordination and Conflict Resolution | 114 |
| 6.3.2 | Authority and Social Alignment | 117 |
| 6.3.3 | User Interaction | 118 |
| 6.3.4 | Information Quality | 119 |
| 6.4 | Wikipedia Article Discussion Corpora | 120 |
| 6.4.1 | Dialog Segmentation | 121 |
| 6.4.2 | Data Sampling | 130 |
| 6.5 | Annotating Wikipedia Article Discussions | 132 |
| 6.5.1 | Annotation Schemes for Article Discussions | 132 |
| 6.5.2 | Corpus Annotation Process and Gold Standard Creation | 137 |
| 6.5.3 | Inter-Annotator Agreement | 139 |
| 6.5.4 | Corpus Analysis | 144 |
| 6.6 | Automatic Prediction of Dialog Act Labels | 146 |
| 6.6.1 | Experiment and System Setup | 146 |
| 6.6.2 | Features | 147 |
| 6.6.3 | Evaluation and Error Analysis | 148 |
| 6.7 | Application Scenario | 151 |
| 6.8 | Chapter Summary | 152 |
| 7 | Summary and Conclusions | 155 |
| 7.1 | Summary | 155 |
| 7.2 | Future Research Directions | 158 |
| | Appendix | 161 |
| A | Open Source Software | 161 |
| B | Annotation Guidelines | 168 |
| C | Cleanup Templates in the English Wikipedia | 180 |
| | List of Tables | 186 |
| | List of Figures | 188 |
| | Bibliography | 203 |
| | Index | 205 |

CHAPTER 1

Introduction

“Begin at the beginning,” the King said gravely, “and go on till you come to the end: then stop.”

— Lewis Carroll, *Alice in Wonderland*

User-generated content is the main driving force of the increasingly social web. Participatory and collaborative content production has largely replaced the traditional ways of information sharing and make up a large part of the daily information consumed by web users. The main properties of user-generated content are a low publication threshold and little or no editorial control. While this has positively affected the variety and timeliness of the available information, it causes an even higher variance in quality than the already heterogeneous quality of traditional web content.

Wikipedia is the prime example for a successful, large scale, collaboratively created resource that reflects the spirit of the open collaborative content creation paradigm. One of the main characters in the popular TV series *The Office* sums up the main idea of Wikipedia in the following ironic quote

“Wikipedia is the best thing ever. Anyone in the world, can write anything they want about any subject. So you know you are getting the best possible information.”¹

In fact, studies ([Giles, 2005](#); [Casebourne et al., 2012](#); [Koistinen, 2013](#)) have confirmed that the overall quality of Wikipedia is high despite its open access policy and lack of rigid regulation. However, there is still a wide gap that must be bridged before Wikipedia reaches the state of a reliable, citable source. In the early days of Wikipedia, the main concern of the community was to increase the coverage of the encyclopedia in order to avoid the problems of its predecessors, which died of insignificance. Today², the English Wikipedia alone

¹Quote from the TV series *The Office*, Series 3, Episode 18.

²Feb 28, 2014

contains as many as 4.5 million articles and the main concern is now “to make Wikipedia as high-quality as possible. [Encyclopædia] Britannica or better quality is the goal.” (LaVallee, 2009)

In order to achieve this goal and provide the “best possible information”, Wikipedia requires a quality management strategy that can cope both with the scale of Wikipedia and its open and almost anarchic nature. Given the fact that less than 10% of Wikipedia users are responsible for more than 90% of the contributions (Ortega, 2009), this quality management strategy cannot rely on the many eyes principle alone to sufficiently assure the quality of all Wikipedia articles. The relatively small core group of active Wikipedians is rather in demand of technical assistance to ensure that even less popular topics in Wikipedia satisfy a basic quality standard.

In this thesis, we discuss how natural language processing approaches can be used to assist the community around Wikipedia in this endeavor. To this end, we consider two basic strategies, a data-driven approach and a process-driven approach. The data-driven approach aims at analyzing and modifying the information directly in order to assess and improve information quality. The latter strategy, on the other hand, aims at improving the established processes involved in analyzing and maintaining the information in order to improve the overall quality of the resource indirectly (Sidi et al., 2012).

Following the data-driven strategy, we present an approach to *automatically identify quality flaws* in Wikipedia articles using state-of-the-art text classification techniques so that passive Wikipedia users, who mainly read articles but are not involved in their maintenance, can be made aware of potential problems in the articles, while active contributors can use this information to solve issues in articles they are interested and experienced in.

As a process-driven strategy, we present an approach to *improve work coordination* between Wikipedians by automatically segmenting and tagging user contributions on article Talk pages, the place where Wikipedia users mainly plan the future development of the articles. These largely unstructured discussion pages are not easy to navigate and information is likely to get lost over time in the discussion archives. By automatically identifying the issues that have been discussed in the past and the solutions that have been proposed, we can *help users to make informed decisions*. The research community can furthermore use this information to *gain insights in the collaborative processes* and identify what differentiates successful work coordination from unsuccessful attempts.

Figure 1.1 shows how the two approaches presented in this thesis relate to each other and to the information quality management process in Wikipedia. In the following section, we give an overview of the contributions of this thesis.

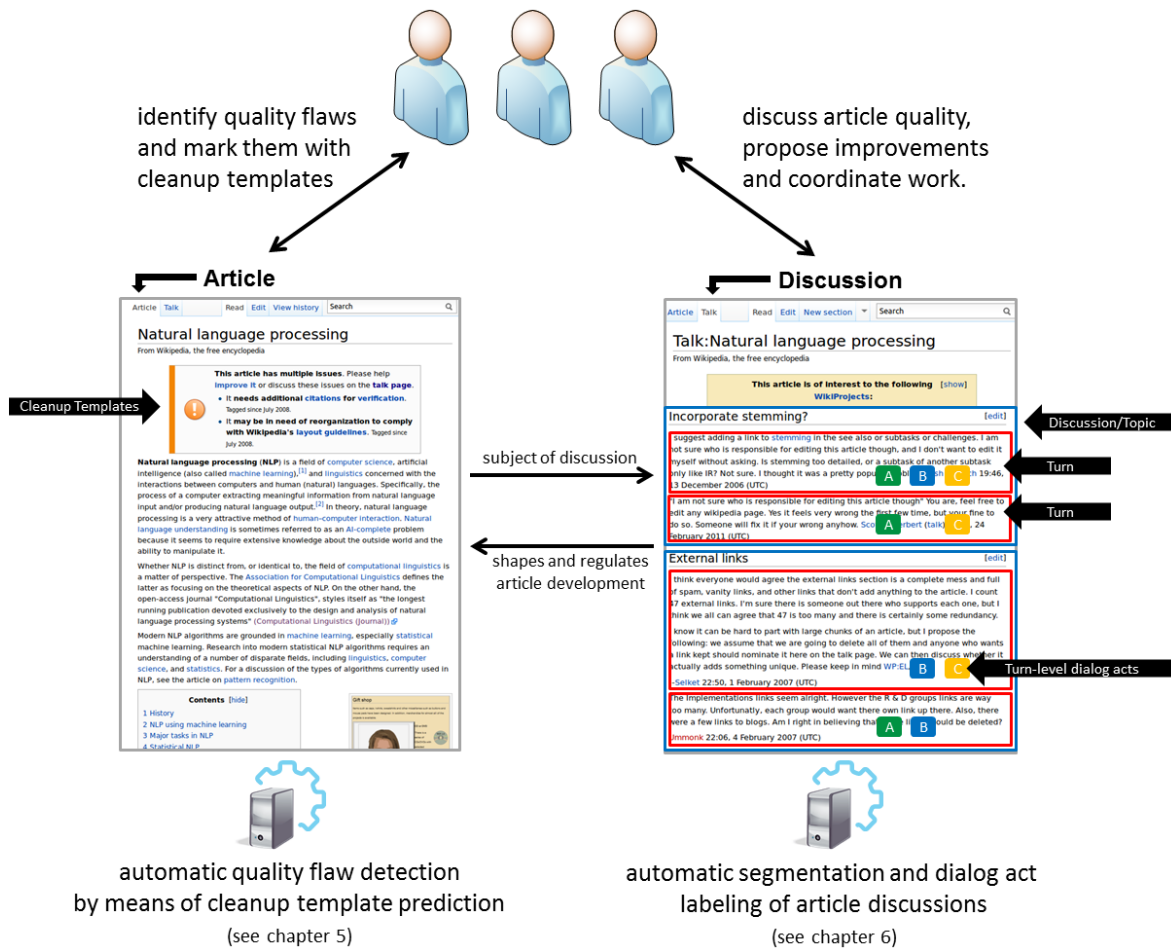


Figure 1.1: Overview of the two approaches to NLP-assisted information quality management in Wikipedia presented in this thesis.

1.1 Main Contributions

The main contributions of this thesis can be divided into a practice-oriented part, i.e. the implications of our work for the information quality management process in Wikipedia, and a theory-oriented part, i.e. the relevance of our contributions to the field of natural language processing. This section gives an overview of our contributions as well as the software and datasets that will be made available to the research community.

- Even though there have been previous attempts to adapt existing information quality frameworks to Wikipedia, no quality model has yet incorporated text and writing quality as a central factor. Since Wikipedia is not only a repository of mere facts but mainly consists of full text articles, the writing quality of these articles has to be taken into consideration when judging article quality. We therefore define a comprehensive article quality model that is based on an information scientific foundation and aims

to consolidate both the quality of writing and the quality criteria defined in multiple Wikipedia guidelines and policies into a single model.

- We present a novel corpus of articles with neutrality and style flaws mined from the English Wikipedia. The corpus contains both articles with the particular flaws and documents that are reliable examples for articles without these flaws. To the best of our knowledge, this is the first corpus of this kind which both provides positive and negative examples for quality flaws.
- For the first time, we establish that automatic quality flaw identification in Wikipedia articles is prone to a topic bias that results in skewed classifiers and unrealistically high cross-validated evaluation results that do not reflect the classifier’s real performance on real world data. We furthermore describe a data sampling approach that is able to avoid this bias in the training data.
- We introduce FlawFinder – a system for supervised text classification designed for quality flaw detection. While FlawFinder has been developed particularly for the flaw detection task, it can be applied to general text classification problems and has been adapted as a general purpose text classification framework that is described in appendix [A.2](#).
- We describe an approach for mining a corpus of quality flaw corrections from Wikipedia’s article revision history which can be used as a starting point for identifying the quality flaws within articles instead of merely tagging whole articles that contain certain flaws.
- We present a novel algorithm for segmenting the unstructured dialog on Wikipedia article Talk pages using the revision history. In contrast to the approaches described in related work which rely on the rudimentary markup and optional user signatures, our algorithm can reliably extract meta information such as the contributor identity and post timestamp even though the information is not contained on the actual page. The algorithm is furthermore able to handle discontinuous turns and inserted replies, which is out of reach from related work.
- We introduce a novel annotation scheme for annotating the turns in article discussions with dialog act labels in order to capture the coordination efforts of article improvement. The labels are intended to reflect the types of criticism discussed in a turn as well as any actions proposed for solving the quality problems. This reflects the core purpose of the discourse on the Wikipedia article Talk pages.
- We present two novel corpora of Wikipedia article discussions extracted from the Simple English Wikipedia (SEWD corpus) and the English Wikipedia (EWD corpus).

The corpora are segmented and manually annotated with dialog act labels defined in our annotation scheme.

- We evaluate how well text classification approaches can learn how to assign the dialog act labels from our scheme to unseen discussion pages. Such classifiers will enable novel applications suitable to improve the work coordination processes in Wikipedia. By automatically identifying the problems that have been discussed and the solutions that have been proposed, we can furthermore gain deeper insights in how the Wikipedia community works and how good work coordination differs from unsuccessful work coordination.

1.2 Publication Record

We have previously published the main contributions of this thesis in peer-reviewed conference or workshop proceedings of major events in natural language processing and related fields, such as ACL, EACL, CLEF and WWW. The chapters which build upon these publications are indicated accordingly. A full bibliography of the author’s publications can be found in the appendix.

Johannes Daxenberger, **Oliver Ferschke**, Iryna Gurevych and Torsten Zesch: ‘DKPro TC: A Java-based Framework for Supervised Learning Experiments on Textual Data’, in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, pp. 61–66, Baltimore, MD, USA, June 2014. (chapters [5](#), [A.2](#))

Lucie Flekova, **Oliver Ferschke** and Iryna Gurevych: ‘What Makes a Good Biography? Multidimensional Quality Analysis Based on Wikipedia Article Feedback Data’, in: *Proceedings of the 23rd International World Wide Web Conference (WWW 2014)*, pp. 855–866, Seoul, Korea, April 2014. (chapter [4](#))

Oliver Ferschke, Iryna Gurevych and Marc Rittberger: ‘The Impact of Topic Bias on Quality Flaw Prediction in Wikipedia’, in: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 721–730, Sofia, Bulgaria, August 2013. (chapter [5](#))

Oliver Ferschke, Johannes Daxenberger and Iryna Gurevych: ‘A Survey of NLP Methods and Resources for Analyzing the Collaborative Writing Process in Wikipedia’, in Iryna Gurevych and Jungi Kim (Edts.): *The People’s Web Meets NLP: Collaboratively Constructed Language Resources*, Chapter 5, pp. 121–160, Springer, April 2013. (chapters [2,3,6](#))

Oliver Ferschke, Iryna Gurevych and Marc Rittberger: ‘FlawFinder: A Modular System for Predicting Quality Flaws in Wikipedia - Notebook for PAN at CLEF 2012’, in: *CLEF 2012*

Labs and Workshop, Notebook Papers, Online Proceedings, Rome, Italy, September 2012. (chapter 5)

Oliver Ferschke, Iryna Gurevych and Yevgen Chebotar. ‘Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages’, in: *Proceedings of the 13th Conference of the European Chapter of the ACL (EACL 2012)*, pp. 777–786, Avignon, France, April 2012. (chapter 6)

Oliver Ferschke, Torsten Zesch and Iryna Gurevych. ‘Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia’s Edit History’, in: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. System Demonstrations*, pp. 97–102, Portland, OR, USA, June 2011. (chapters 3,5,6,A.1)

1.3 Thesis Organization

In the remainder of this chapter, we give an overview of the organization of this dissertation.

Chapter 2 discusses the foundations of open collaboration and introduces the main characteristics of collaborative and open collaborative writing. It furthermore gives a brief overview of systems for joint online writing and how they can benefit from language technology.

Chapter 3 introduces the free online encyclopedia Wikipedia and defines the terminology necessary to perform the succeeding analyses. We introduce both the main characteristics of the encyclopedia and the culture and community that emerged around it. Additionally, we introduce technological aspects such as the different possibilities to process the content of Wikipedia.

Chapter 4 discusses the theory of information quality and how it applies to a collaboratively created resource such as Wikipedia. We discuss the processes involved in defining an information quality model and finally adapt an established, generic model to the specific needs of Wikipedia under particular consideration of text and writing quality.

Chapter 5 presents our novel approach for automatically identifying quality flaws in Wikipedia based on supervised detection of cleanup templates. These templates are assigned to articles by Wikipedia users and identify concrete shortcomings of an article. We argue that these markers are suitable proxies for quality flaws and, in turn, an adequate means for quality assessment and the basis for assisting users in quality improvement. Moreover, we identify a methodological problem in existing approaches for quality flaw

detection and establish that the flaw prediction task inherently suffers from a topic bias which has to be accounted for in any machine learning attempt. We describe a solution to this problem in our approach and present a novel corpus of neutrality and style flaws that is already controlled for the topic bias.

Chapter 6 introduces our approach to analyzing Wikipedia article discussion pages by means of dialog act analysis. We develop a scheme for identifying user contributions discussing quality problems and suggesting actions to solve these problems. We present two annotated corpora extracted from the Simple English Wikipedia and the English Wikipedia respectively. We furthermore carry out text classification experiments on both corpora in order to evaluate how well a machine learning algorithm can automatically apply labels to unseen turns in order to automate the dialog act analysis task.

Chapter 7 draws conclusions from the preceding chapters and summarizes both the solved problems and the challenges that still remain to be addressed in future work.

1.4 Terminology and Conventions

Unless otherwise specified, any references to Wikipedia and Wikipedia content refer to the English Wikipedia. Whenever any specific content on a wiki page is referenced or cited, we provide the revision or access date of the corresponding page, e.g. <http://en.wikipedia.org/wiki/index.php?oldid=596488753>. In cases where a policy or guideline page is referenced as a concept without referring to the specific content on this page, we provide the shortcut to the page, e.g. <http://en.wikipedia.org/wiki/WP:MOS> for the Wikipedia Manual of Style.

CHAPTER 2

Open Online Collaboration

“Individually, we are one drop. Together, we are an ocean.”

— Ryunosuke Satoro

In this chapter, we discuss the properties of open collaboration with a focus on collaborative writing. We first identify the main characteristics of this fairly new *modus operandi* in joint work and analyze the important factors for implementing a suitable information quality management strategy with language technology assistance (section 2.1). We then discuss collaborative writing as a special instance of collaboration, how it differs from individual writing and how open online collaboration adds additional levels of complexity to the writing task (section 2.2). We finally provide a brief discussion regarding the requirements of systems for collaborative online writing (section 2.3) and conclude the chapter with a summary of our findings (section 2.4).

2.1 Open Online Collaboration

The term *computer supported cooperative work* (CSCW) exists since 1984 when it was coined by Irene Greif and Paul Cashman as the title of a workshop on understanding and supporting collaboration (Grudin and Poltrock, 2013). While, in the beginning, CSCW has mainly focused on the utilization of computer-mediated communication, such as email, to support online collaboration in research- or corporate workgroups, the field soon transcended these realms and now increasingly penetrates other aspects of our daily social and work interactions. Most notably, with the rise of the general availability of the Internet, CSCW has expanded from the confines of small groups to open, heterogeneous communities.

Closed group vs. Open Collaboration. While online collaboration in closed groups often relies on fixed role and task assignments and is frequently steered by a higher authority,

open collaboration is an egalitarian and meritocratic process in which everyone who joins the group can contribute to the best of their ability to an iteratively improving product, while the merits of each contribution can be publicly discussed by the community. Instead of assigned tasks and fixed roles, open collaboration is self-organizing to a high degree with the objective that every collaborator can find their own mode of participation and perform tasks they are interested in and qualified for.

The difference between closed group and open collaboration is well illustrated by two metaphors originated in the context of open software development (Raymond, 1999). The *cathedral model* compares the collaborative process with the construction of a cathedral that has to be planned in advance, supervised by experts and carried out by a fixed group of contractors who collaborate to build a single, final product – the cathedral. This resembles traditional closed-group collaboration, which is often directed at a particular final goal which is to be reached with a predefined plan and fixed task assignments. If any collaborator fails, the whole project is endangered.

In contrast, the *bazaar model* compares the collaborative process to a bazaar on which many people trade their goods without being controlled by a central authority. Each marketer has equal opportunities, equal rights and is able to choose the individual contribution to the community on their own. The bazaar as a whole is complete and functional even with individual stalls and merchants being absent. This resembles open collaboration, which is an inherently iterative process without a fixed workflow or static role assignments. The product organically evolves as a result of swarm creativity, i.e. the aggregated individual contributions of the changing set of collaborators, and does not necessarily reach a final state of a finished product but rather remains in constant evolution.

Applications. The range of applications for open online collaboration is wide and often closely connected to the concept of social networks (Forte and Lampe, 2013). Platforms embracing the Web 2.0 spirit attract a large crowd of users whose workforce is put to joint use for the collaborative creation of online maps (e.g. OpenStreetMap³), news (e.g. Slashdot⁴, Digg⁵), dictionaries (e.g. Wiktionary⁶) or encyclopedias (e.g. Wikipedia⁷), just to name a few examples.

Challenges in Open Collaboration. According to Forte et al. (2012), self-organizing communities have to face three main challenges. First, since participation in open collaboration is usually intrinsically motivated, *incentives and motivation* play essential roles in achieving long-term success. New users have to be attracted while current members of the

³<http://www.openstreetmap.org>

⁴<http://www.slashdot.org>

⁵<http://www.digg.com>

⁶<http://www.wiktionary.org>

⁷<http://www.wikipedia.org>

community must be retained and kept motivated to actively contribute to the collaborative project. Second, it is necessary to *bring the people to the work* instead of counting on any individual to self-select the ideal job. In other words, the available work force has to be distributed and allocated to open tasks. It is necessary for the whole community to be aware at all times which open issues have to be addressed and where the greatest demand for contributions is. Since no central management exists in open collaboration, this has to be achieved by means of collaborative *work coordination*. Third, since centralized decision making is unfeasible in very large heterogeneous groups, sub-communities with nested organizational structures have to emerge which concentrate on particular sets of tasks. These sub-communities develop their own social dynamics and thereby help to maintain the morale and trust among their members. Together, the output of the individual sub-communities contributes to the overall product of the open community. In Wikipedia, for example, this is mainly achieved with so-called WikiProjects, sub-communities that focus on particular subject areas or maintenance tasks.

Universal Properties. Even though no two open collaboration communities are alike and each project has different goals, one can identify universal properties of open collaboration. The *power law of participation* implies a long tail of many collaborators with few contributions and little impact on the system while a small group of elite users is responsible for the main body of work. Consequently, the overall collaborative system and the product it produces are mainly shaped by the small groups of experts (Ortega et al., 2008). It furthermore has to be taken into account that the reasons for participation in open collaboration are different for each individual and that these reasons shape their level of activity and the type of work they do. It is therefore not sufficient to attract many users but it is rather important to attract enough users for every critical task (Forte and Lampe, 2013).

Collaboration Support Systems. While open collaboration is not necessarily confined to the realm of the world wide web, online cooperation is the most common mode of open collaboration. A key factor for its success is the support by a suitable online collaboration system that assists the open community in their endeavors. Forte and Lampe (2013) identify four dimensions of socio-technical systems for open online collaboration that have to be taken into account. In short, an open online collaboration system has to provide assistance for the collective production of an artifact, it has to support the social aspects of collaboration and work coordination by providing suitable means of communication, it has to reduce the complexity in order to lower the entry barrier for new collaborators and has to assist the development and upkeep of social structures in order to retain the active population of the community.

Quality Management. In open, collaborative content production, a key element of success is the quality of the content produced by the community. Therefore, a suitable quality management strategy is an absolute requirement for the success of such a community. We will address the issue of information quality management in the context of the collaboratively created encyclopedia Wikipedia in chapter 4 and discuss in the succeeding chapter how language technology can assist the process.

2.2 Collaborative Writing

In the narrowest sense, *writing* is the externalization of natural language in a visual or tactile form based on a formalized writing system. Rather than looking at the mechanics of writing, we are interested in the intellectual process of text production, including all related activities such as brainstorming, idea generation, planning, organizing, drafting and revising (Rice and Huguley J.T., 1994). In the latter sense, writing is an open-ended design task without a single correct end result that could be reached in a clearly defined chain of operations. It rather is a creative process that involves nondeterministic sequences of edits, revisions, deletions and amendments which finally lead to one of many possible texts that are suitable solutions for the given writing task (Sharples et al., 1993).

Traditionally, writing is thought of as a process involving a single author who iteratively plans, drafts and reviews his work (Lowry et al., 2004). In order to cope with large-scale writing tasks in a limited amount of time, authors have, however, always collaborated with each other. Collaborative writing resembles individual writing in many ways but is inherently more complex on a social, intellectual and procedural level (Galegher and Kraut, 1994). While writing is ultimately a process of externalizing thoughts and ideas, a large part of the writing process takes place in the mind. Collaborating with other writers in the joint goal to create a single, collaboratively created text therefore inevitably requires externalizing the otherwise hidden thoughts during the writing process for the sake of work coordination. However, despite this added complexity, research on collaborative writing has shown that often the result of good collaboration is more than the sum of its parts (Sharples, 1993).

Several disjunct theories and models for collaborative writing have been developed across the fields, each with their own terminology. In an attempt to build an interdisciplinary taxonomy and nomenclature of collaborative writing, Lowry et al. (2004) identifies the common aspects of collaborative writing across different fields of research and defines four basic strategies of joint writing:

Group Single-Author Writing: *A single author compiles the results of a collaborative planning phase. The writing process itself is largely self-directed thus resembling more the process of individual writing than collaborative writing. (figure 2.1a)*

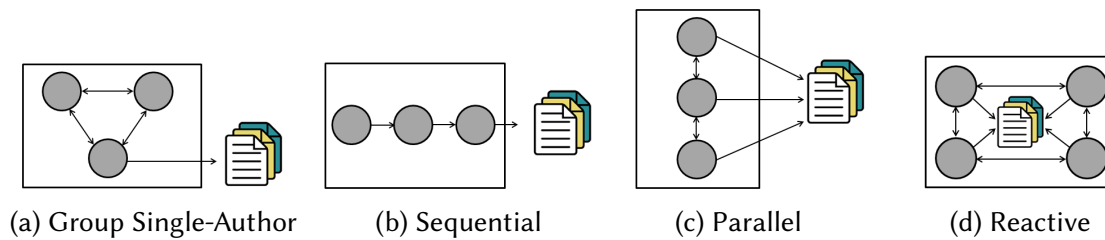


Figure 2.1: Collaborative writing strategies according to [Lowry et al. \(2004\)](#)

Sequential Writing: *An iterative writing strategy in which the text is passed on from group member to group member. Only one author is writing at the same time. The required work coordination during the active writing phase is minimized, while the overall planning process and the contributions of the group members can be distributed. On the downside, social interaction can be limited in this approach. (figure 2.1b)*

Parallel Writing: *In this strategy, all group members simultaneously work on the same document. Parallel writing can be divided into horizontal division writing and stratified division writing. In the former case, each member of the group works on a separate sub-document which is finally merged into the final document. In the latter case, each member takes a different role in the writing process, such as editor, author and reviewer, and processes the document from a different point of view. (figure 2.1c)*

Reactive Writing: *The most challenging strategy involves concurrent editing of the same document by all group members. This approach is only viable with suitable support by the collaboration system and is the main strategy in open online collaboration. It demands the highest degree of coordination. (figure 2.1d)*

While the collaborative aspect in *group single-author writing* is restricted to the planning phase of writing, *sequential writing* involves the incorporation of different writing styles into a single document. However, sequential writing is only efficient for small groups and not suitable for larger documents. The output volume of this strategy is furthermore limited because only a single group member is allowed to actively contribute to the document at a given point in time. In *parallel writing*, these problems are solved by distributing the writing task across all group members allowing everyone to edit simultaneously but in different sections of the text. This increases the requirements for coordination but makes the task more efficient at the same time. However, the explicit assignment of authors to sections of the text also restricts the strategy to small groups. Even though *reactive writing* is technically the most challenging strategy with the highest requirements for coordination, it is the only approach that scales to web size, i.e. is suitable for large-scale open online collaboration. Any group member can directly edit any part of a document giving them the opportunity to decide for themselves how and what to contribute to the final product. We therefore consider this strategy to be the only suitable approach for open online collabo-

rative writing. Depending on the work mode of the collaboration system, group members either edit in real-time (live editing) or in near real-time (asynchronous, revision-based editing).

Collaborative writing, or any kind of writing for that matter, is not a single, atomic activity but rather a complex process made up of different interlocked sub-activities. These activities have to be properly supported by the collaboration system in order to facilitate the joint writing process. [Lowry et al. \(2004\)](#) define seven activities of collaborative writing:

Brainstorming: *Development of new ideas by all group members.*

Convergence on brainstorming: *Ranking and filtering of the ideas developed in the brainstorming phase.*

Outlining: *Organization of ideas in a high level outline that sketches the rough structure of the document.*

Drafting: *Filling the outline with content in order to create a first incomplete draft of the text.*

Reviewing: *Adding comments and suggestions for corrections to the draft.*

Revising: *Responding to the comments from the review phase in order to create an improved draft.*

Copyediting: *Making final corrections to the draft in order to create a final, consistent document.*

These activities should not be seen as strictly sequential. They can rather be combined in an arbitrary order and be revisited in any stage of the writing process. While it is possible for any group member to participate in every phase of the writing process, it is often the case that individual contributors specialize in particular tasks, i.e. they take different roles in the writing process such as author, copy-editor or reviewer. While the role assignment is strictly pre-assigned in writing strategies such as stratified-division writing (see above), open forms of collaborative writing will leave it up to the contributors to self-select for individual roles.

Open Collaborative Writing As we have established before, collaborative writing adds additional complexity to the writing process compared to individual writing mainly due to the added coordination efforts necessary to synchronize the work of all co-authors. Open collaboration in writing adds yet another level of complexity to the process since the coordination strategies have to scale to a large, heterogeneous group with different levels of expertise and different agendas.

Open collaborative writing is still a relatively rare phenomenon compared to other efforts of joint work that we already listed above. Many efforts never exceeded the state of exploratory experiments. [Rettberg \(2005\)](#), for example, lists several activities revolving around the idea of *constructive narratives* following the concepts of *exploratory* and

constructive hypertext coined by Joyce (1988). Collaborative writing projects such as the *Hypertext Hotel*⁸, *The Unknown*⁹ or *1001 Nights Cast*¹⁰ resemble long term writing events producing a *work in progress* which is intended to remain in constant evolution rather than becoming a single finished product.

Some of the main lessons learned in these experiments was that it is important to limit the extent of openness in open collaboration in order to reach satisfactory results. That is, in order to synchronize the efforts of all participants, they have to agree upon constraints that everyone has to abide to and come to explicit agreements regarding their cooperation. Without coordination and agreement, the collaborative efforts are prone to incoherence and previous work is likely to be reverted by later contributors.

This holds not only true for literary experiments like the ones listed above, but also for more down-to-earth projects with real life relevance such as Wikipedia. Even though the participation in Wikipedia is *open* and not regulated by predefined, fixed rules, it is necessary to coordinate the work of many in order to reach a common goal – a high quality encyclopedia. Wikipedia is introduced in detail in chapter 3, while the issues of quality management in the context of open collaboration will be discussed in chapter 4.

2.3 Collaborative Online Writing Systems

According to Lowry et al. (2004, p. 92), a *collaborative writing system* is a piece of “[s]oftware that allows collaborative writing groups to produce a shared document and assist collaborative writing groups to perform the major collaborative writing tasks.” This subsumes a wide range of tools ranging from mere version control to full-blown writing environments.

Noël and Robert (2004) carried out an empirical study interviewing 33 individuals in a web survey about the most important aspects of collaborative writing tools. Among the highest ranked answers, the participants mentioned synchronous access to the documents, version control, communication between collaborators, ability to add stand-off comments to the text, visualization of the version history and spaces to plan and schedule the future work on the documents. Even though there was a high variation in the answers, the above mentioned aspects occurred multiple times stressing their importance in different application scenarios.

It is beyond the scope of this work to define a hierarchy of collaborative writing systems and weight the pros and cons of each possible incarnation. Noël and Robert (2003) give a detailed overview of 19 web-based systems for collaborative writing and discuss the merits and problems of each solution. While Noël and Robert exclusively focus on asynchronous collaboration system, many recent tools foster synchronous collaboration, i.e. they allow

⁸<http://netlern.net/hyperdis/hyphotel> accessed on Feb 27, 2014

⁹<http://unknownhypertext.com> accessed on Feb 27, 2014

¹⁰<http://1001.net.au> accessed on Feb 27, 2014

users to collaborate in real time on the same document. Examples for such tools are *Ether-Pad*¹¹, *Google Docs*¹² or *Zoho Docs*¹³, but also traditional word processing software, such as *Microsoft Word*¹⁴, expand their scope into the web in order to support real-time collaboration.

Wikis. Even though not originally designed as a writing tool, the *wiki* has particularly taken hold as a collaborative writing system. The term *wiki* stems from the Hawaiian expression *wiki wiki*, which translates to “very fast” and illustrates the main focus of the technology – fast and easy content production and management with minimal overhead (Leuf and Cunningham, 2001).

Wikis are web-based, asynchronous co-authoring tools whose content is structured with lightweight markup that is translated into HTML by the wiki system. The markup is restricted to a small set of keywords, which lowers the entrance barrier for new users and reduces the barrier to participation. Many wiki systems even offer visual editors that automatically produce the desired page layout without having to know the markup language.

A unique characteristic of wikis is the automatic documentation of the revision history keeping track of every change that is made to a wiki page which can also be visually represented. With this information, it is possible to reconstruct the writing process from the beginning to the end and revert malicious changes in order to restore an earlier, clean version of the document. Additionally, many wikis offer their users a communication platform, the *Talk pages*, where they can discuss the ongoing writing process with other users.

Thus, wikis satisfy all the above listed requirements for good collaborative writing platforms. In the course of this thesis, we focus on *Wikipedia*, a wiki-based encyclopedia, which is one of the most successful collaborative online projects on the world wide web.

Chances for NLP assistance. While the openness of wikis, their low entry barrier and the lightweight markup are the main reasons for their success, they are, at the same time, the major points of concerns for large-scale projects that aim at high quality content. As wikis and their user base grow, they tend to become unstructured and unorganized (Buffa, 2006).

Recent research has suggested to use NLP to automatically improve the structure of the wikis and transform them into self-organizing content management systems while retaining openness and ease of use and without imparting too many restrictions on the users (Hoffart et al., 2009; Bär et al., 2011). While these efforts are mainly directed towards improving the usability of intranet wikis that are used as knowledge management systems and

¹¹<http://etherpad.org>

¹²<https://docs.google.com>

¹³<https://www.zoho.com/docs>

¹⁴*Microsoft Office 365* is a subscription-based online software that offers, among others, collaborative word processing and spreadsheet capabilities.

thus rather resemble the closed group collaboration paradigm, *large-scale, open* wikis aimed at collaborative content production, such as Wikipedia, rather demand a quality management strategy that informs the users at all times of any quality problems and demand for improvement while offering an effective communication platform on which the work can be coordinated. NLP can offer improvements in both of these aspects. Using the example of Wikipedia, this thesis will present two approaches to improve information quality management at large scale in open collaborative environments.

2.4 Chapter Summary

In this chapter, we have discussed the foundations of open collaboration. We identified the key differences between closed group and open collaboration and discussed the main challenges involved in the latter, such as motivation, coordination and work allocation. We further discussed universal properties of open collaboration that can be found across all collaborative platforms. In particular, the power law of participation suggests an unequal work distribution across all users while the motivation of each user to contribute to the collaborative project differs and thus influences the nature of their contribution.

In a second part of the chapter, we turned to collaborative writing and analyzed how it differs from individual writing. We discussed the main collaborative writing strategies and identified typical writing activities involved in order to inform any decisions that aim at improving the work coordination and quality management in a collaborative environment.

We finally addressed open collaborative writing, which adds another level of complexity to the collaborative writing task. We established that it is necessary to reduce the complexity of the task by explicitly constraining the openness with policies and guidelines upon which all participants have to agree. This is not done top-down from a central authority, but by coordination of all users. Therefore, effective means for work coordination and policy making are necessary.

We closed the chapter by reviewing the typical requirements for collaborative writing systems and identified the wiki as a system that satisfies the most important requirements. We furthermore suggested that NLP can help to overcome the inherent problems coming along with the openness of these systems, their low entry barrier and lack of a central, regulatory authority.

CHAPTER 3

Wikipedia

“Wikipedia is the best thing ever. Anyone in the world, can write anything they want about any subject. So you know you are getting the best possible information.”

— Michael Scott (played by Steve Carell), *The Office*

This chapter aims at introducing Wikipedia and defining the basic terminology. We first give a general overview of the online encyclopedia and its evolution (section 3.1) and introduce its main structure and organization (section 3.2). We proceed with a discussion of the community around Wikipedia, the understanding of which is vital for any quality-related analysis (section 3.3). We then examine Wikipedia’s versioning system – the revision history (section 3.4) – and its communication hub – the discussion pages (section 3.5). We furthermore discuss technical aspects of automatically processing the large amount of data Wikipedia has to offer (section 3.6). We conclude the chapter by introducing the most important sister-projects of Wikipedia (section 3.7) and summarize our findings (section 3.8).

3.1 Overview

As the previous chapter has shown, wikis have proven to be a suitable and well-accepted technology for large-scale online collaboration. The most prominent example of a successful, large-scale wiki is *Wikipedia*, a free, collaboratively created online encyclopedia which is available in 287 languages and dialects¹⁵. Its main website, *wikipedia.org*, ranks in the TOP 10 of the most visited pages on the web according to the Alexa web traffic report¹⁶. While the term *Wikipedia* usually refers to this website, it also describes the community behind the encyclopedia that plans, discusses and creates its content.

¹⁵According to http://meta.wikimedia.org/wiki/List_of_Wikipedias as of 6 Sept 2013

¹⁶Rank 7 according to <http://www.alexa.com/siteinfo/wikipedia.org> as of 2 Sept 2013

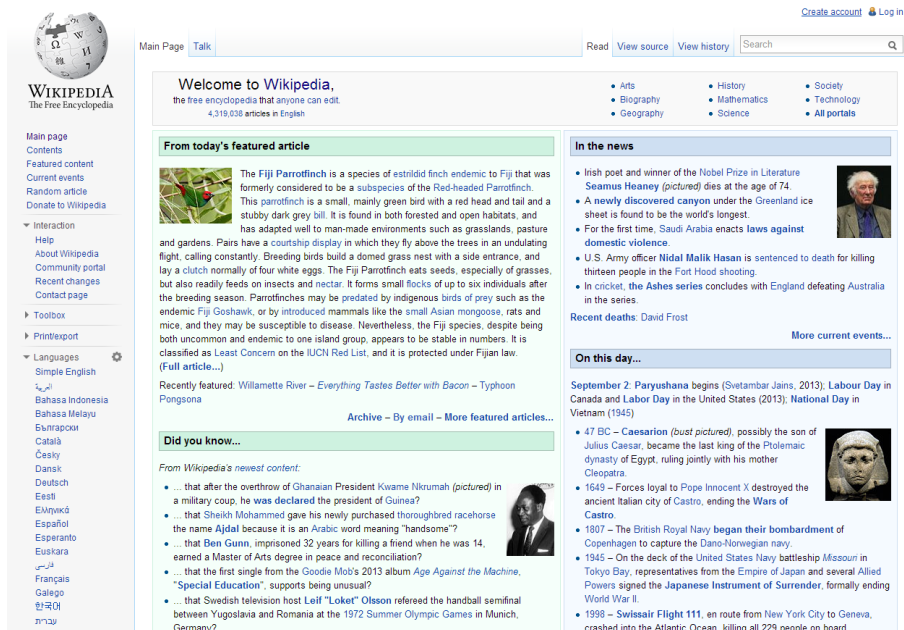


Figure 3.1: Main page of the English Wikipedia
<http://en.wikipedia.org> on 2 Sept 2013

History. The idea of creating a collaborative online encyclopedia goes back to 1993 when Rick Gates first proposed his project *Interpedia* which, however, never left the planning stage. The idea was picked up again in 2000 by Jimmy Wales who started *Nupedia* as a free online encyclopedia. Unlike its successor Wikipedia, Nupedia was not an open collaborative platform, but relied solely on expert content and employed a complex seven-layer review system. This restrictive policy, however, caused Nupedia to attract little attention both by readers and contributors. In an attempt to attract a wider audience, Wales established Wikipedia as a side project and, at the same time, an incubator for Nupedia articles. Founded on January 15 2001, Wikipedia quickly became popular by word-of-mouth resulting in 1,000 articles to be created within the first year. Nupedia could never step out from under Wikipedia’s shadow and was closed in 2003 with only 24 completed articles (Ayers et al., 2008; Reagle, 2010).

United in diversity. Even though Wikipedia is often referred to as a *multilingual encyclopedia*, it is more precisely described as an encyclopedia that comes in many interlinked language versions. While the former definition implies a uniformity in the organizational and administrative structures across all languages, the latter definition better captures the individual nature and culture of each language community. Each language version of Wikipedia has its unique governance policies, quality standards and perception of what may and may not be included in the encyclopedia. These regulations have naturally grown over the years in a collaborative attempt to find a common ground within the language commu-

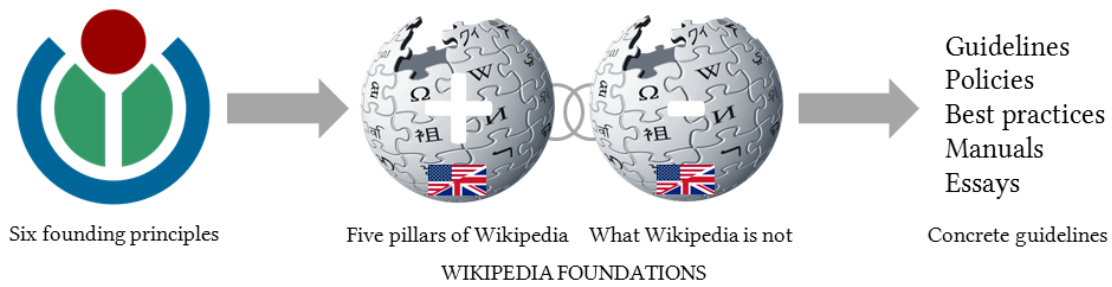


Figure 3.2: The origin of policies in Wikipedia using the example of the English Wikipedia

nity that best serves the needs of both the creators and consumers. Despite all differences, there is a set of six *founding principles*¹⁷ which are shared by all Wikimedia projects (see section 3.7 for an overview) and, by extension, all language editions of Wikipedia: These generic principles have been adapted to reflect the peculiarities of an encyclopedia resulting in the so-called five pillars of Wikipedia¹⁸

1. Wikipedia is an encyclopedia.
2. Wikipedia is written from a neutral point of view.
3. Wikipedia is free content that anyone can edit, use, modify, and distribute.
4. Editors should treat each other with respect and civility.
5. Wikipedia does not have firm rules.

This specific interpretation of the six founding principles is shared by most language editions of Wikipedia with only minor deviations. In particular, the fifth pillar has often been subject of debate and is not accepted by some language versions, such as the German Wikipedia¹⁹. In addition to defining the main characteristics of Wikipedia, most language versions also provide a set of *demarcation criteria* which define the limits of Wikipedia²⁰. Together, they form the *Wikipedia Foundations* and represent the basis for all further policies and practical guidelines (see figure 3.2)

- Wikipedia is not a paper encyclopedia.
- Wikipedia is not a dictionary.
- Wikipedia is not a publisher of original thought.
- Wikipedia is not a soapbox or means of promotion.
- Wikipedia is not a mirror or a repository of links, images, or media files.
- Wikipedia is not a blog, Web hosting service, or social networking service.
- Wikipedia is not a directory.
- Wikipedia is not a manual, guidebook, textbook, or scientific journal.

¹⁷http://meta.wikimedia.org/wiki/Founding_principles/de

¹⁸<http://en.wikipedia.org/wiki/WP:5>

¹⁹<http://de.wikipedia.org/w/index.php?oldid=122405554> accessed on 10 Sept 2013

²⁰<http://en.wikipedia.org/wiki/WP:NOT>

- Wikipedia is not a crystal ball.
- Wikipedia is not a newspaper.
- Wikipedia is not an indiscriminate collection of information.
- Wikipedia is not censored.

Within the confines of the Wikipedia Foundations, all language communities establish their own cultures and regulations. This inevitably results in the development of different philosophies regarding key aspects such as article organization²¹, notability standards²², article development²³ and various other issues²⁴. Even though these phenomena can also be observed within any larger Wikipedia, they are most evident when comparing different language versions.

Despite the pursuit of a neutral point of view as one of the founding principles, the different cultural backgrounds of the individual language communities inevitably introduce point of view differences across the Wikipedias. A seemingly neutral and objective article may exhibit a notable cultural bias that describes the subject matter in a more positive or negative light than the same article in a different language version of Wikipedia (Massa and Scrinzi, 2011; Al Khatib et al., 2012). This so-called *systemic bias*²⁵ is particularly evident in larger Wikipedias with users from different cultural backgrounds, such as the English Wikipedia, because the cultural diversity of the community is a prerequisite for becoming aware of this problem. Less culturally diverse Wikipedias might suffer from the same problem which, however, may remain undetected by the community (also see section 3.3.3).

3.2 Structure and Organization

As a compromise between the low publication threshold of the wiki concept and the structural requirements of a large-scale encyclopedia, Wikipedia takes the middle ground between an unstructured and a semi-structured resource by exhibiting traits of both worlds. While the content in Wikipedia is strongly interconnected with different types of links and redirects, contains structured elements, such as infoboxes, and makes use of a sophisticated category system providing a high degree of ontologization, Wikipedia still relies on a low entrance barrier which makes it possible for new users to contribute without any particular training (Hovy et al., 2013).

In the remainder of this section we will first discuss the macrostructure of Wikipedia, i.e. the inter-page organization, and then proceed with the microstructure, i.e. the intra-page organization.

²¹Lumpers vs. Splitters: http://en.wikipedia.org/wiki/Lumpers_and_splitters

²²Deletionists vs. Inclusionists: http://en.wikipedia.org/wiki/Deletionism_and_inclusionism

²³Eventualists vs. Immediatists: Rettberg (2005)

²⁴http://meta.wikimedia.org/wiki/Conflicting_Wikipedia_philosophies

²⁵http://en.wikipedia.org/wiki/Wikipedia:Systemic_bias

| Namespace | | English | French | German | Spanish | Russian |
|-----------|---|------------------|-----------------|-----------------|-----------------|-----------------|
| Main | A | 10,454,480 (58%) | 2,708,931 (47%) | 2,719,812 (41%) | 2,491,636 (59%) | 2,110,411 (50%) |
| | T | 5,213,744 (11%) | 1,242,508 (10%) | 573,230 | 232,007 (1%) | 412,143 |
| User | A | 1,740,863 (7%) | 201,732 (4%) | 364,470 (4%) | 140,336 (4%) | 91,181 (4%) |
| | T | 9,057,248 | 1,216,634 | 379,595 (1%) | 1,125,806 | 295,688 |
| Wikipedia | A | 793,364 (13%) | 36,257 (23%) | 40,303 (11%) | 24,759 (9%) | 33,801 (9%) |
| | T | 210,651 (25%) | 5,517 (26%) | 11,187 (14%) | 2,009 (13%) | 2,006 (4%) |
| File | A | 843,156 | 41,654 | 172,730 | 0 ^a | 155,219 |
| | T | 175,705 | 1,582 | 2,173 | 0 ^a | 958 |
| MediaWiki | A | 1,895 (1%) | 1,027 | 1,786 | 1,508 | 988 |
| | T | 1,081 (11%) | 139 (14%) | 230 (12%) | 94 (2%) | 144 (3%) |
| Template | A | 524,227 (19%) | 162,243 (7%) | 54,526 (2%) | 19,573 (19%) | 100,493 (14%) |
| | T | 204,315 (12%) | 6,230 (7%) | 3,865 (5%) | 1,132 (2%) | 5,581 (2%) |
| Help | A | 1,327 (52%) | 931 (45%) | 854 (71%) | 240 (45%) | 25 (72%) |
| | T | 629 (24%) | 391 (26%) | 375 (22%) | 77 (9%) | 5 |
| Category | A | 1,053,314 (2%) | 246,800 | 185,706 | 215,871 | 246,182 |
| | T | 676,963 | 14,962 | 6,621 | 3,008 | 4,553 |
| Portal | A | 120,635 (8%) | 49,445 (10%) | 16,518 (7%) | 12,465 (9%) | 19,219 (3%) |
| | T | 30,069 (5%) | 3,339 (41%) | 4,566 (16%) | 511 (6%) | 722 (1%) |
| Other | A | 5,007 (10%) | 121 (2%) | 69 | 51 | 198 |
| | T | 4,705 (8%) | 27 (4%) | 26 (92%) | 16 (6%) | 6 |
| Total | | 31,113,117 (23%) | 5,940,443 (24%) | 4,538,616 (25%) | 4,271,083 (35%) | 3,479,517 (31%) |

^a No pages in this namespace. This Wikipedia exclusively embeds media information from Wikimedia commons.

Table 3.1: Number of pages per namespace for the five largest Wikipedias as of 4 Sept 2013. The percentage of redirects is provided in parentheses, if applicable. *A* denotes article namespaces, *T* the corresponding talk namespaces.

3.2.1 Namespaces and Naming Conventions

While the best known artifacts in Wikipedia are the articles, which contain the encyclopedic content, a substantial fraction of all pages serve administrative and communicative purposes. Wikipedia is organized in so-called namespaces, a system of thematic layers which group pages according to their main purpose. In addition to the main encyclopedic layer, there are, for example, namespaces that hold administrative pages, help pages, user pages and descriptions of media assets. Each of these namespaces has an associated Talk namespace, which holds discussion pages related to the corresponding content pages (see section 3.5). Overall, eight default subject namespaces are predefined by the MediaWiki software²⁶:

Main: *Contains encyclopedic articles, lists, disambiguation pages and redirects.*

User: *Contains user pages and sub-pages created by individual users. This namespace is often used as an incubator for new content in the main namespace.*

²⁶<http://www.mediawiki.org/wiki/Manual:Namespace>

Project: *The project namespace contains pages about the wiki-project itself. In the case of Wikipedia, this project namespace is also named Wikipedia. It contains policy pages, best practices, workflows and essays about the work in Wikipedia.*

File: *Contains pages with descriptions of media items including links to these media items. The actual media items are hosted on the Wikimedia Commons platform²⁷. Pages in this namespace are only used to override the original media descriptions on Wikimedia commons.*

MediaWiki: *Contains internal content provided by the MediaWiki installation, such as standard system messages.*

Template: *Contains template pages that can be inserted into other pages (see section 3.2.4)*

Help: *Contains help pages for passive and active users of Wikipedia.*

Category: *Contains pages for every category which list the members of this category along with an optional category description.*

In addition to these default namespaces, Wikipedia defines custom namespaces that hold pages for Wikipedia-specific features such as thematic portals or Wikipedia book projects. These namespaces might vary across the different language versions. Table 3.1 shows statistics of page numbers per namespace for the five largest Wikipedias.

Like the entries of most traditional encyclopedias, Wikipedia articles correspond to single concepts. The article naming conventions²⁸ hereby ensure that the titles are recognizable, natural, precise, concise and consistent with titles of similar articles. Since natural language tends to be ambiguous, it is necessary to handle polysemous page titles, i.e. titles that may refer to multiple concepts. In Wikipedia, this is achieved by means of natural disambiguation, comma-separated disambiguation or, in most cases, parenthetical disambiguation.

Natural Disambiguation: *An alternative, non-ambiguous title is used that also meets the naming conventions. This is the preferred disambiguation form.*

Example: *Instead of English, use English language or English people*

Comma-separated Disambiguation: *The disambiguation term is added as a comma-separated suffix to the title, if it stands in a hierarchical relationship to the main concept. It is most commonly used with geographic names.*

Example: *Lincoln, Nebraska ; Lincoln, England ; Lincoln, New Hampshire*

Parenthetical Disambiguation: *The disambiguation term is added as a parenthetical suffix to the title. This is the most common disambiguation form.*

Example: *Apple (fruit) ; Apple (computer) ; Apple (album)*

In order to establish a link between disambiguated terms and their polysemous lemma, the latter is used as the title of a disambiguation page. A disambiguation page lists all senses of a

²⁷<https://commons.wikimedia.org>

²⁸<http://en.wikipedia.org/wiki/WP:NAMINGCRITERIA>

polysemous term which are represented in Wikipedia and links to their respective articles. These pages are internally flagged as disambiguation pages, since they do not count as content pages.

Another type of pages without content are redirects. Redirect pages are used to represent multiple variants of page titles that refer to the same concept. This way synonymous terms, writing variants (e.g. British English vs. American English) and different words forms (e.g. verb inflections) can be mapped to a single page. If a concept in Wikipedia can be described by several terms that equally qualify as page titles, the most salient term is chosen as the title of the content page for this concept, while the other possible terms are used as titles for redirect pages. Redirect pages simply forward to the corresponding content page and are marked as such in the database. Table 3.1 lists the percentage of redirects in each namespace in parentheses.

3.2.2 Organizational Structures

In order to facilitate the navigation through the encyclopedia beyond the full text search, Wikipedia provides several organizational structures such as categories, lists and portals. As we will show later, the proper use of these structures has a great impact on the overall quality of the encyclopedia, since not only the quality of the content is important, but also its organization.

Category System. The *Wikipedia category system* is the most comprehensive organizational structure in Wikipedia, comprising over a million categories in the English Wikipedia as of 4 September 2013 with four different category types:

Administrative categories: *Indicate the maintenance status of articles*

Example: “*Articles needing cleanup*”

Container categories: *Group other categories, but do not directly apply to articles*

Example: “*People categories by parameter*”

Set categories: *Represent lists of articles*

Example: “*Cities in Germany*”

Topic categories: *Group articles related to a particular topic*

Example: “*History of Germany*”

Even though the category system is organized hierarchically, it is rather a thematic classification used for tagging wiki pages than a well-defined taxonomy for document categorization (Nagata et al., 2010; Syed and Finin, 2010). The hierarchical relationships between the categories form a large graph structure, the *Wikipedia Category Graph* (WCG), which was found to be a scale-free, small-world graph similar to lexico-semantic networks such as WordNet (Zesch and Gurevych, 2007).

Lists. Besides the extensive category system, Wikipedia offers other means of organizing articles. *Lists* are pages that link thematically related articles. In contrast to the *set category*, which also serves the purpose of grouping related articles, lists can also include links to non-existing articles (so-called *redlinks*) as reminders that these articles still have to be written. Lists are created manually and are subject to similar quality standards as articles.²⁹ Like categories, lists are organized hierarchically on up to three levels (see *Lists of lists of lists*³⁰). Since these pages are maintained by hand, they are prone to inconsistency and incompleteness.

In addition to generic lists, Wikipedia offers several other devices for content organization which resemble lists to a large extent. These include *glossaries* for listing term definitions, *outlines* and *overviews* for hierarchically organizing links to the main articles on particular topics, and *indexes*, which provide alphabetical, automatically created overviews of all articles in particular subject areas.

Portals. In contrast to categories and lists, *portals* do not merely provide links to articles, they rather serve as an entry point to the main topics in Wikipedia. They provide the reader with excerpts of well selected articles from each subject area and even include links to related external resources or news items related to the subject. Portals are often associated with and maintained by a *WikiProject* (see section 3.3), which manages and coordinates the article development in a specific subject area. The Wikipedia main page (see figure 3.1) can be regarded as a special portal for the purpose of bringing distinguished or particularly noteworthy content to the attention of the reader.

3.2.3 Inner Article Structure

While no particular article layout is enforced by the wiki software, the Wikipedia guidelines demand a basic common article structure (Ayers et al., 2008). Figure 3.3 shows an overview of an example article and its building blocks. The body of the article is usually segmented into titled sections. The titles of these sections are used by the wiki software to automatically create a table of contents for the article. The introductory or lead paragraphs sum up the content of the article in a short, self-contained text and give an overview of the scope of the article. The optional *infobox* on the top of the page is a fixed-format table which summarizes basic facts about the article subject or provides links to related articles. These infoboxes are one of the few sources of structured information in Wikipedia. Throughout the article text, links to other wiki pages, so-called *wikilinks*, are provided in order to interconnect related pages and to provide definitions for concepts that are not explained in

²⁹The criteria for distinguished lists can be found under <http://en.wikipedia.org/wiki/WP:FLCR>. Similar criteria exist for articles, as will be discussed in detail in chapter 4.

³⁰http://en.wikipedia.org/wiki/List_of_lists_of_lists

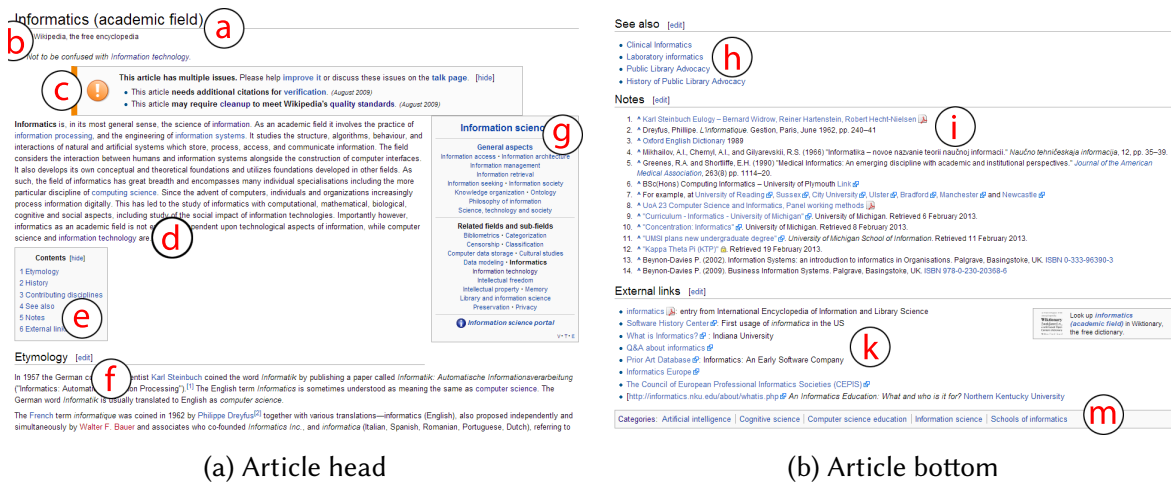


Figure 3.3: Structure of an article: *a*) article title with parenthetical disambiguation, *b*) hatnote, *c*) warning messages, *d*) introductory/lead paragraph, *e*) table of contents, *f*) first section with section title, *g*) infobox, *h*) links to other articles, *i*) bibliography, *k*) links to external resources, *m*) category memberships

Source of example: <http://en.wikipedia.org/wiki/index.php?oldid=584601250>

the article itself. Links colored in red (*redlinks*) point to articles that do not yet exist and serve as a reminder to create these articles. On the top of the page, so-called *hatnotes* display important information such as disambiguation terms or redirects. Optional warning messages on the top of the page furthermore inform the reader of existing problems with the article or ongoing debates. They are produced by templates, which are discussed in the following section. The bottom of each page displays footnotes and references and provides the bibliography for the article. Furthermore, it contains links to external resources, since these are not allowed to be placed directly in the article text. Every article closes with a list of all categories of which the article is a member.

Outside of the main article frame, in the left navigation bar, *interwiki links* are listed, which link to corresponding articles in other language versions of Wikipedia. As of now, these links have to be maintained separately in each article of every language version. However, the *Wikidata* project currently attempts to develop a centralized system for maintaining these interwiki connections (see section 3.7).

Articles are written and formatted in *wiki markup*, a lightweight markup language that is designed to hide the complexity of richer markup languages, such as (X)HTML, from the user. This way, using wiki markup is intended to be as simple as using natural language. However, with the rising demand for complex article layouts and integrated functions for automation purposes, the simplicity of the wiki markup language could no longer be maintained. While the core of the markup language is the same as in the early days, countless additions to the instruction set have made it difficult to use for new users and even impos-

sible to parse reliably for computers.³¹ The latter issue, the irregularity and ambiguities of the wiki markup language, is the reason why no WYSIWYG editor has been developed to date that is fit to serve the purposes of the Wikipedia community. In an ongoing large-scale project, the Wikimedia Foundation develops the *Visual Editor*³², which is supposed to be able to reliably parse and generate wiki markup and thus enable WYSIWYG editing of Wiki pages. This is believed to further lower the entry boundary for new Wikipedia contributors and to lower the time necessary to edit an article.

3.2.4 Template System

In principle, *templates* are small wiki pages that can be embedded in another page in order to centralize repetitive content. A common use case for templates is to embed info banners, system messages, warnings or navigational boxes into articles or other wiki pages.

Depending on the particular template, the embedded content is either transcluded (i.e., inserted into the page on runtime but not in the source code) or substituted (i.e., inserted directly in the source code). The content of templates is usually not static, but it can be controlled with a set of parameters passed to the template in the wiki markup. For example, the structure of an infobox can be centrally defined in a template while the actual content is defined in the embedding articles. This way, the information provided by a certain type of infobox is uniform across all articles using it. The *Wikidata* project (see section 3.7) further attempts to centralize the data for infoboxes across different language versions of the same articles in order to improve the overall consistency of the encyclopedia.

While parameters are often used to inject messages or data into the template, like in the case of infoboxes, the range of possibilities is much wider. Via the MediaWiki extension *Scribunto*, it is possible to include scripts written in the *Lua*³³ programming language directly in the source code of the template pages. These scripts can be controlled with the parameters that are passed to the templates and produce the content that is finally displayed on the embedding page.

Besides including recurring content or embedding output of Lua-scripts, templates are frequently used as a tagging system. A prominent example for this usage are *cleanup templates*, which aim at identifying articles with particular deficiencies. Whenever a cleanup template is embedded on a wiki page, it both displays a message on the page and adds the page to the corresponding cleanup category. This way, it is easy to keep track of the problems marked by the templates via the category system while the problems are at the same time communicated to the readers via the embedded message. In chapter 5, we use these cleanup templates as human assigned labels to create corpora of quality flaws in Wikipedia.

³¹http://www.mediawiki.org/wiki/Markup_spec

³²<http://en.wikipedia.org/wiki/WP:VE>

³³<http://www.lua.org>

3.3 Community

Even though we mainly regard Wikipedia as a collaboratively created resource, it is not only an encyclopedia but also a community. In order to understand how the resource is constructed and how it evolves, one has to obtain a basic understanding of the community behind the encyclopedia.

3.3.1 User Groups and Roles

While the five pillars of Wikipedia state that anyone can edit the encyclopedia, different user groups define who is or is not allowed to perform particular actions in Wikipedia and furthermore manage the responsibilities and competences within the self-administration of the community. The following list describes the major user status groups in Wikipedia³⁴:

Unregistered User: *Any user who is not registered or not logged in is identified with their IP address. Depending on the protection status of certain pages, unregistered users might not be able to perform simple edits.*

Registered User: *Any logged-in user who is registered with a valid email address. Most non-privileged actions are available for this status group.*

Reviewer: *Trusted user who is allowed to review edits of other users made to articles under the pending changes protection or with flagged revisions. (see section 3.4)*

Administrator: *Users with increased edit privileges who can delete and restore pages, block users, protect pages, manage some user groups and perform other maintenance functions. The administrator role is local to a certain Wikipedia language version.*

Bureaucrat: *Users with privileges necessary for user right management such as user group assignment or change of user names.*

Ombudsman: *A small group of users with increased access rights who investigate violations of privacy policies across Wikimedia projects.*

Steward: *A small group of users with complete access to all Wikimedia wikis, including the ability to change user rights and groups.*

Developer: *Users with the highest degree of technical access, since they are able to directly make changes to the MediaWiki software and the underlying data.*

Even though anonymity is an important aspect for many members of the Wikipedia community, accountability still has to be ensured. Therefore, editing anonymously as an unregistered user is frowned upon while registered users usually use pseudonyms which protect

³⁴A full list of additional user groups along with a detailed description of the respective access rights can be found under http://en.wikipedia.org/wiki/http://en.wikipedia.org/wiki/Wikipedia:User_access_levels for the English Wikipedia and under https://meta.wikimedia.org/wiki/User_groups for general Wikimedia projects.

their privacy but still ensure accountability of their actions (Ayers et al., 2008). Furthermore, it is expected that every registered individual only participates with a single account. Multiple accounts owned by a single user are called *sock puppets*, which might be used to create the illusion of greater support or rejection of a particular issue and hence influence the collaborative decision making process towards the goals of the user. Automatically detecting sock puppets can be regarded as an instance of authorship attribution and can contribute to the quality management process (Solorio et al., 2013).

3.3.2 Soft Security

Even though the available user groups provide some security by limiting general access to non-registered users with technical means, the open collaboration principle still allows most changes to be made by any registered user. Rather than imposing rigid restrictions on users willing to contribute to the project, Wikipedia follows the approach of *soft security*. By assuming good faith in every (anonymous) participant and relying on the many eyes principle, it is assumed that high quality output can be reached and any damages made to the resource in the course of the collaborative process can be kept within tolerable limits. Furthermore, with a transparent and open administrative system, everyone can be included in the decisions that govern the collaborative process (Ayers et al., 2008; Reagle, 2010). Even though this approach has been very successful for Wikipedia so far, some Wikipedias have now reached a size where peer review and the many eyes principle alone can no longer assure the integrity and quality of the whole resource without any technological assistance, which remains to be shown in the course of this thesis.

3.3.3 Systemic Bias

Since Wikipedia articles are the collaborative product of the Wikipedia community, they naturally represent the views and values of this community. The principle of open participation and the international availability of Wikipedia suggests that the community is a balanced representation of the world's population. However, studies have shown that this is not the case (Lam et al., 2011; Glott et al., 2010; Hill and Shaw, 2013). The majority of Wikipedia authors and active community members are white males from western countries resulting in a gender gap and a western-centric world view. This *systemic bias* is in conflict with Wikipedia's policy of the neutral point of view, but is one of the hardest issues to resolve.

3.3.4 WikiProjects

*WikiProjects*³⁵ are communities of interest aimed at providing a forum for contributors who work together on a particular subject area or provide a particular service for the community such as article maintenance. As of October 2013, there exist over 2,200 WikiProjects in the English Wikipedia³⁶, which are centrally registered in the WikiProject Directory³⁷.

While each WikiProject has to obey the general Wikipedia guidelines, larger groups establish their own policies, quality management workflows and quality standards. Since WikiProjects and their members are considered to be most knowledgeable about their particular subject area, they are responsible to provide quality feedback for the articles in their field and decide about inclusion and exclusion of particular topics. Even though the quality standards might differ across WikiProjects, they all make use of the same labels indicating the same quality levels. This way, it is possible to aggregate the quality information centrally to gain a general overview of the quality of Wikipedia at a given point in time (see chapter 4.3).

3.4 Revision History

A key characteristic of Wikipedia is its revision history which keeps track of all changes that have ever been made to any wiki page. Every time a page is edited by a Wikipedia user, a new version of this page is created. We call each individual version of a Wikipedia page a *revision*, denoted as r_v . v is a number between 0 and n , where r_0 is the first and r_n the newest version of the page. In addition to the full text of the page with markup, the Wiki system stores for each revision additional metadata, such as the user who changed the page and thus created the new revision, the creation time of the revision, an optional commit comment and a flag whether minor or major changes have been made to the page (also see section 3.6 for a detailed overview of all information stored in the database). One of the main goals of this versioning system is to provide the possibility to revert the changes made to a page in one or more revisions and thus return to an earlier content state of the page. Each revert will again result in a new revision of the reverted page.

Since every single revision is self-contained and the text of the page is stored in full for each revision, the content of the revision history is highly redundant. This drastically increases the amount of space necessary to store the data and results in large data dumps (see section 3.6). In addition to revisions, we define *diffs* to be the set of all changes between two revisions while each individual, atomic change is called an *edit*. A single diff can therefore comprise one or multiple edits. The MediaWiki allows to display diffs in

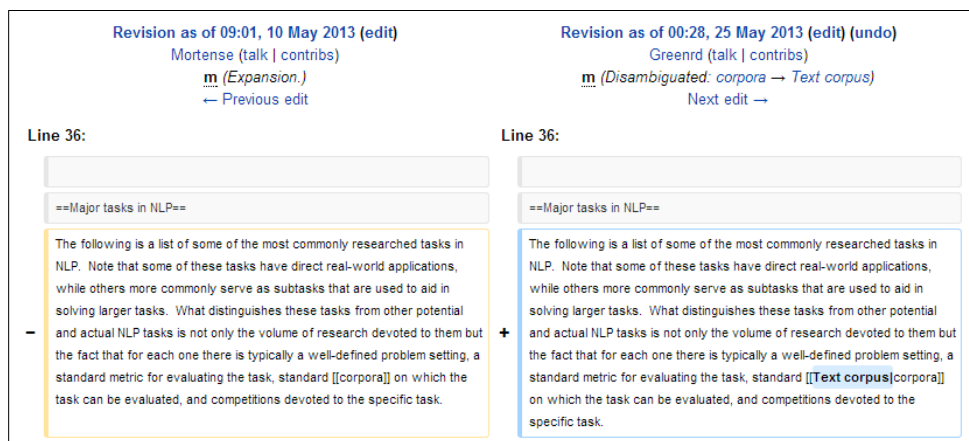
³⁵<http://en.wikipedia.org/wiki/WP:PROJ>

³⁶<http://en.wikipedia.org/wiki/index.php?oldid=576344994>

³⁷<http://en.wikipedia.org/wiki/WP:PROJDIR>



(a) Excerpt of a revision history



(b) Diff between two revisions displayed on a DiffPage in the MediaWiki software

Figure 3.4: Revisions of the article “Natural Language Processing” in the English Wikipedia accessed on 05.Jan 2014

DiffPages, highlighting all changes identified in a line-by-line comparison of two revisions. Figure 3.4 shows an excerpt of the revision history as it is presented in the MediaWiki software along with a DiffPage comparing two adjacent revisions of an article.

Privileged users can limit page changes to users of particular status groups, for instance, editing can be restricted to registered users or users with particular privileges, such as administrators. These restrictions can either prevent particular types of edits, such as renaming a page, or can apply to any kinds of changes. A detailed description of the available protection levels can be found in the *Wikipedia Protection Policies*³⁸.

In rare cases, for instance if confidential information has been provided on a page that might violate the privacy of an individual, particular revisions or the revision history of whole pages can be deleted by privileged users in order to prevent any access by the public.

³⁸<http://en.wikipedia.org/wiki/WP:PP>

There have been efforts to introduce quality control mechanisms on the revision level in form of so-called *flagged revisions*³⁹, which allow experienced users to moderate the edit activities of new users. This editorial review was intended to minimize the risk of vandalism and improve the accuracy and overall quality of the articles by having experienced Wikipedia authors approve revisions before they go public. While this approach was accepted by the community of the German Wikipedia very early and is being used successfully as part of the information quality management process, other Wikipedia communities display mixed sentiment regarding the system. The English Wikipedia now uses a modified version of the flagged revisions called *pending changes*⁴⁰, which currently only requires edits of unregistered and newly registered users to be reviewed.

While the benefit of maintaining a revision history is obvious for the users of Wikipedia, it also serves as an invaluable resource for natural language processing applications. We provide a detailed list of related work based on the Wikipedia revision history in [Ferschke et al. \(2013\)](#).

3.5 User Discussions

As we have discussed in chapter 2, authors of collaboratively written texts have to externalize processes that remain hidden in individual writing, such as the planning and organization of the text. Work coordination is particularly important in open collaboration, since explicit workflows which regulate the writing process do not exist and individual users might have different goals regarding the further development of an article.

In Wikipedia, the main platform for work coordination and user communication are the Talk pages. Technically speaking, a Talk page is a normal wiki page located in one of the Talk namespaces (see table 3.1 and table 3.2). Similar to a web forum, Talk pages are divided into discussions (or topics) and contributions (or turns). What distinguishes wiki discussions from a regular web forum, however, is the lack of a fixed, rigid thread structure. There are no dedicated formatting devices for structuring the Talk pages besides the regular wiki markup. The basic structure of an article Talk page can be seen in the example shown in figure 3.5).

Each Talk page is implicitly connected to a content page by its page name—e.g. the Talk page `Talk:Germany` corresponds to the article `Germany`. It is, however, not possible to establish explicit connections between individual discussions on the page and the section of the article that is being discussed. Each namespace in Wikipedia has a corresponding Talk

³⁹<http://en.wikipedia.org/wiki/WP:FLR>

⁴⁰<http://en.wikipedia.org/wiki/WP:PC>

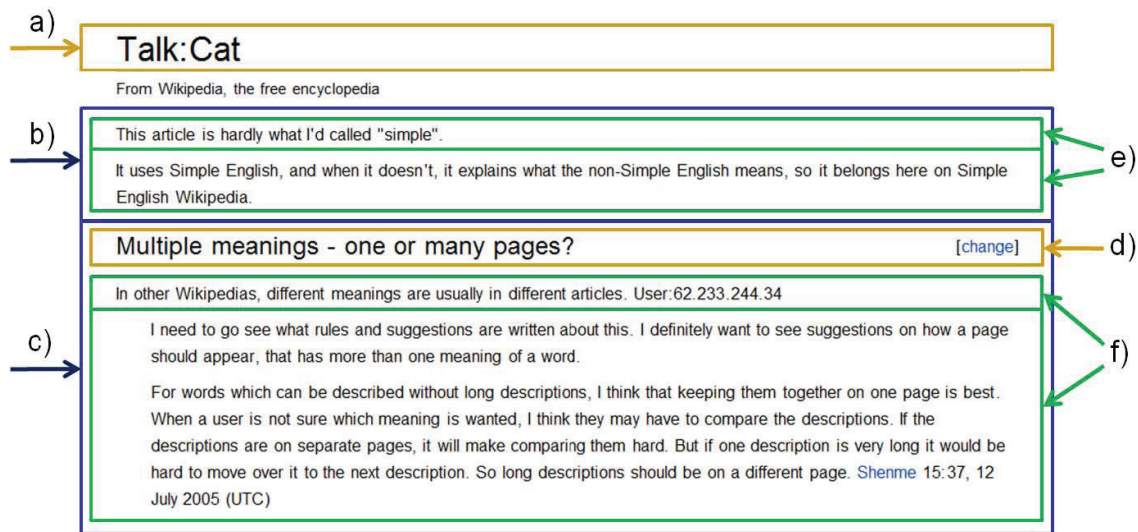


Figure 3.5: Structure of a Talk page: *a)* Talk page title, *b)* untitled discussion topic, *c)* titled discussion topic, *d)* topic title, *e)* unsigned turns, *f)* signed turns

Source of example: <http://simple.wikipedia.org/w/index.php?oldid=4633184>

Visualization first appeared in Ferschke et al. (2012a)

namespace resulting in a total of ten⁴¹ major types of Talk pages in the English Wikipedia (table 3.2) which can be categorized into four functional classes:

Article Talk pages are mainly used for the coordination and planning of articles.

User Talk pages are used as the main communication channel and social networking platform for the Wikipedians.

Meta Talk pages serve as a platform for policy making and technical support.

Item-specific Talk pages are dedicated to the discussion of individual media items (e.g. pictures) or structural devices (e.g. categories and templates).

The users are asked to structure their contributions using paragraphs and indentation. One *turn* may consist of one or more paragraphs, but no paragraph may span over several turns. Turns that reply to another contribution are supposed to be indented to simulate a thread structure. We call this *soft threading* as opposed to *explicit threading* in web forums.

Users are furthermore encouraged to append signatures to their contributions to indicate the end of a turn (see figure 3.6). There are extensive policies⁴² that govern the usage and format of signatures. They usually should contain the username of the author and the

⁴¹While there are additional special purpose namespaces which have been subsumed under *Other* in table 3.2.1, we only list the *Book* namespace here, since it is the only one with significant discussion activity. Other minor namespaces in the English Wikipedia include *Education Program*, *TimedText* and *Module*.

⁴²<http://en.wikipedia.org/wiki/WP:SIGNATURE>

| Content namespaces | Talk namespaces | Functional class |
|--------------------|-----------------|------------------|
| Main | Talk | Article |
| User | User talk | User |
| Wikipedia | Wikipedia talk | Meta |
| MediaWiki | MediaWiki talk | Meta |
| Help | Help talk | Meta |
| File | File talk | Item |
| Template | Template talk | Item |
| Category | Category talk | Item |
| Portal | Portal talk | Item |
| Book | Book talk | Item |

Table 3.2: Wikipedia namespaces and functional Talk page classes

time and date of the contribution. However, users’ signatures do not adhere to a uniform format, which makes reliable parsing of user signatures a complex task. Moreover, less than 70% of all users explicitly sign their posts (Viégas et al., 2007). In some cases, depending on the setup of an individual Talk page, automatic scripts—so-called “bots”—take over whenever an unsigned comment is posted to a Talk page and add the missing signature (see figure 3.6, signature 3.5). While this is helpful for signature-based discourse segmentation, which relies on the presence of a user signature to identify turn boundaries, it is misleading when it comes to author identification where the actual content of the signature is important.

Due to the lack of discussion-specific markup, contribution boundaries are not always clear-cut. They may even change over time, for instance if users insert their own comments into existing contributions of other users, which results in non-linear discussions. This makes automatic segmentation of Talk pages a challenging task and demands a substantial amount of preprocessing. We will again refer to this phenomenon under the term *in-text replies* when discussing our own approach for dialog segmentation in chapter 6.4.1.

There are ongoing attempts to improve the usability of the discussion spaces with extensions for explicit threading⁴³ and visual editing⁴⁴. However, these enhancements have been tested only in selected, small Wikimedia projects and have not yet been deployed to the larger wikis.

In order to prevent individual Talk pages from becoming too long and disorganized, individual discussions can be moved to a discussion archive⁴⁵. A general policy states that Talk pages with more than ten discussion topics or a size of more than 75 Kilobytes should be archived. However, the requirements differ depending on the discussion activity of a given Talk page. Discussion archives are marked with an “Archive” suffix and usually num-

⁴³<http://www.mediawiki.org/wiki/Extension:LiquidThreads>

⁴⁴<http://en.wikipedia.org/wiki/WP:VE>

⁴⁵<http://en.wikipedia.org/wiki/WP:ARCHIVE>

- The Rambling Man (talk) 18:20, 27 February 2012 (UTC) (3.1)
- 66.53.136.85 21:41, 2004 Aug 3 (UTC) (3.2)
- Taku (3.3)
- Preceding unsigned comment added by 121.54.2.122 (talk) 05:33, 10 February 2012 (UTC) (3.4)
- SineBot (talk) 08:43, 31 August 2009 (UTC) (3.5)
- Imzadi 1979 > 09:20, 20 May 2011 (UTC) (3.6)
- ♪Greatorangepumpkin♪ 14:14, 17 December 2010 (UTC) (3.7)

Figure 3.6: Examples for user signatures on Talk pages: (3.1) Standard signature with username, link to user Talk page and timestamp (3.2) Signature of an anonymous user (3.3) Simple signature without timestamp (3.4,3.5) Bot-generated signatures (3.6,3.7) Signatures using colors and special Unicode characters as design elements

bered consecutively. The oldest discussion archive page for the article “Germany”, for example, is named `Talk:Germany/Archive_1`. There are two possible procedures for archiving a Talk page: the *cut-and-paste procedure* and the *move procedure*. While it is not possible to determine directly which method has been used to create an archive, the choice has important implications for processing these pages. The cut-and-paste procedure copies the text from an existing Talk page to a newly created archive page. All revisions of this Talk page remain in the revision history of the original page. The move procedure renames (i.e., moves) an existing Talk page and adds the numbered archive suffix to its page title. Afterwards, a new Talk page is created that is then used as the new active Talk space. Archives created with the latter procedure maintain their own revision history, which simplifies the revision-based processing of these pages.

Furthermore, in addition to automatic archiving, topic specific sub-pages might be created for particularly focused discussion, e.g. the discussion of a request for deletion (RFD) or the review process involved when an article is nominated for promotion to featured or good article status (see chapter 4.3).

Although there is no discussion-specific markup to structure Talk pages, *templates* (see section 3.2.4) can be used to better organize the discussions. A specific subset of templates is used as a tagset for labeling articles and Talk pages. For example, by adding the template `{{controversial}}` to a Talk page, an information banner is placed in the lead section of the Talk page and the associated article is tagged as controversial. A complete overview of Talk space specific templates can be found on the corresponding Wikipedia policy pages⁴⁶. The

⁴⁶<http://en.wikipedia.org/wiki/WP:TTALK>

cleanup and flaw markers are especially helpful criteria for filtering articles and Talk pages for corpus creation or further analysis.

In chapter 6, we focus on article Talk pages and discuss how they are employed for work coordination and how we can exploit them as a resource for computational linguistics. While the different kinds of Talk pages are the main communication platform on Wikipedia, some aspects are discussed outside of the confines of the MediaWiki software and communicated via mailing lists⁴⁷, Internet Relay Chat (IRC) channels⁴⁸ or real-life meetings⁴⁹.

3.6 Processing Wikipedia

Wikipedia runs on the MediaWiki wiki software⁵⁰, which offers multiple possibilities for accessing its contents. Depending on the requirements of the application at hand, be it structured access to particular pieces of information or large-scale processing of the whole data contained in Wikipedia, several options are available, which we discuss in this section. First, we describe the different available sources for Wikipedia data, then we discuss APIs, software libraries and services which can be used to access these data sources. A more detailed description of the software developed in the course of this thesis can be found in appendix A.

3.6.1 Data Sources

The MediaWiki software stores all content and meta information in an SQL database with over fifty tables⁵¹. Except for sensitive user information and some other privileged artifacts, this database can be fully accessed in various ways, as the next subsection will show. For many purposes, however, offline images of the data are needed, which represent a fixed snapshot of the whole resource and which can be processed locally.

In order to provide the highest degree of interoperability, XML dumps of the MySQL databases are provided for download⁵². Partial dumps can furthermore be created manually via the export function of the MediaWiki software⁵³. The information included in these dumps is described in the document type definition of the XML format, which is listed

⁴⁷<https://lists.wikimedia.org>

⁴⁸<http://en.wikipedia.org/wiki/WP:CHAT>

⁴⁹Regular face to face meetings in smaller interest groups (<http://en.wikipedia.org/wiki/WP:MEET>) are as well organized as larger, non-academic conventions for users of Wikimedia projects, such as *WikiMania* (<http://wikimania2014.wikimedia.org>) or *WikiCon* (<http://de.wikipedia.org/wiki/WP:CON>).

⁵⁰<http://www.mediawiki.org>

⁵¹http://www.mediawiki.org/wiki/Manual:Database_layout

⁵²<http://dumps.wikimedia.org>

⁵³<http://en.wikipedia.org/wiki/Special:Export>

```

1 <!ELEMENT mediawiki (siteinfo ,page *)>
2 <!ATTLIST mediawiki
3   version CDATA #REQUIRED
4   xmlns CDATA #FIXED "http://www.mediawiki.org/xml/export-0.3/"
5   xmlns:xsi CDATA #FIXED "http://www.w3.org/2001/XMLSchema-instance"
6   xsi:schemaLocation CDATA #FIXED "http://www.mediawiki.org/xml/export-0.3.xsd"
7   xml:lang CDATA #IMPLIED
8 >
9 <!ELEMENT siteinfo (sitename ,base ,generator ,case ,namespaces)>
10 <!ELEMENT sitename (#PCDATA)> <!-- name of the wiki -->
11 <!ELEMENT base (#PCDATA)> <!-- url of the main page -->
12 <!ELEMENT generator (#PCDATA)> <!-- MediaWiki version -->
13 <!ELEMENT case (#PCDATA)> <!-- how cases in page names are handled -->
14 <!ELEMENT namespaces (namespace+)> <!-- list of namespaces and prefixes -->
15 <!ELEMENT namespace (#PCDATA)> <!-- contains namespace prefix -->
16 <!ATTLIST namespace key CDATA #REQUIRED> <!-- internal namespace number -->
17 <!ELEMENT page (title ,id?,restrictions?,(revision|upload)*)>
18 <!ELEMENT title (#PCDATA)> <!-- title with namespace prefix -->
19 <!ELEMENT id (#PCDATA)> <!-- unique id of page -->
20 <!ELEMENT restrictions (#PCDATA)> <!-- optional page restrictions -->
21 <!ELEMENT revision (id?,timestamp ,contributor ,minor?,comment?,text)>
22 <!ELEMENT timestamp (#PCDATA)> <!-- according to ISO8601 -->
23 <!ELEMENT minor EMPTY> <!-- minor revision flag -->
24 <!ELEMENT comment (#PCDATA)> <!-- commit comment -->
25 <!ELEMENT text (#PCDATA)> <!-- wiki markup -->
26 <!ATTLIST text xml:space CDATA #FIXED "preserve">
27 <!ELEMENT contributor ((username ,id) | ip)>
28 <!ELEMENT username (#PCDATA)> <!-- username of contributor -->
29 <!ELEMENT ip (#PCDATA)> <!-- ip of contributor -->
30 <!ELEMENT upload (timestamp ,contributor ,comment?,filename ,src ,size)>
31 <!ELEMENT filename (#PCDATA)> <!-- name of uploaded file -->
32 <!ELEMENT src (#PCDATA)> <!-- location of uploaded file-->
33 <!ELEMENT size (#PCDATA)> <!-- size of uploaded file -->

```

Figure 3.7: Document Type Definition (DTD) for the MediaWiki export format describing the content of a Wikipedia dump. Adapted from <http://meta.wikimedia.org/wiki/Help:Export>, accessed on 20 Dec 2013.

in figure 3.7. Dumps of large Wikipedia language versions including page revisions are very large in size, since every revision is self-contained and contains the full page text (see section 3.4). As of 2013, the decompressed XML dump for the English Wikipedia exceeds eight terabytes. In order to handle amounts of data of this size, the dumps are split into several, independent XML files and individually compressed.

While the XML dumps contain all page contents of Wikipedia, more volatile information, such as page view statistics, user group assignments and page ratings, can be downloaded as additional SQL files.

| Name | Functionality | API | License |
|-----------------|------------------------------------|-------------|---------|
| MediaWiki API | access all publicly available data | web service | – |
| Wikimedia Labs | access all publicly available data | database | – |
| JWPL | access articles and Talk pages | Java | LGPL |
| WRT | access page revisions | Java | LGPL |
| Wikipedia Miner | access articles | Java | GPL |
| WikiXRay | quantitative statistics | Python, R | GPL |
| WikiHadoop | process Wikipedia dumps in Hadoop | Java | ASL |

Table 3.3: Tools and services for accessing Wikipedia. References are provided in section 3.6.2

3.6.2 Data Access

This section gives an overview of different tools and service for accessing the information in Wikipedia. We do not include small, special purpose scripts provided by Wikipedia users⁵⁴, but rather restrict the discussion to general purpose APIs, software libraries and services (see table 3.3).

The MediaWikiAPI provides direct access to the databases of the MediaWiki installations which underly each Wikimedia project.⁵⁵ It is available as a web service⁵⁶ with wrappers for various programming languages⁵⁷. While the API supports many queries, provides various output formats and delivers up-to-date information, it is not suitable for most moderate- and large-scale processing tasks, since the performance of the web service is very limited and poses severe restrictions on non-privileged users. The API is used mostly by bots (maintenance scripts), which receive privileged access rights in order to perform maintenance activities within Wikimedia projects. In contrast to most other access tools, the MediaWikiAPI cannot only be used to retrieve data, but also to update and add data.

The Wikimedia Labs is a scalable, cloud-based test and development environment that provides virtual machines on which individual code can be run with direct access to replicated versions of the live Wikimedia databases, including all language versions of Wikipedia and Wiktionary.⁵⁸ As the successor project of the *Wikimedia Toolserver*⁵⁹, the *Wikimedia Labs* aim at providing both a development environment and a place for hosting online tools intended to be used directly by the community.

⁵⁴A compilation of these can be found under http://en.wikipedia.org/wiki/WP:WikiProject_User_scripts/Scripts

⁵⁵<http://www.mediawiki.org/wiki/API>

⁵⁶<http://en.wikipedia.org/w/api.php>

⁵⁷http://www.mediawiki.org/wiki/API:Client_code

⁵⁸<https://wikitech.wikimedia.org>

⁵⁹<http://toolserver.org>

The unique advantage of running software in the Labs environment is the direct database access. While not granting access to restricted information, these databases offer all information that is publicly available for each Wikimedia project. This includes a wider range of information than the downloadable XML data dumps have to offer. Furthermore, the databases are always up-to-date. Beyond that, no particular API is provided to access the data in a structured manner.

Depending on the software requirements, code can either be hosted on an individual virtual machine, or it can be deployed to the Tool Labs, which aggregate smaller tools related to Wikimedia projects. This offers a new way of disseminating applications which originate in research projects for use by the wider public.

As of the time of writing, no long-term experiences with the Wikimedia Labs have been reported regarding the performance of the runtime environment, because the project is still in an early beta phase. The former Toolserver struggled with performance issues, which made large scale processing of Wikipedia data infeasible. However, this is supposed to be solved in the Wikimedia Labs.

The Java Wikipedia Library (JWPL) [Zesch et al. \(2008\)](#) offer a Java-based programming interface for accessing all information in different language versions of Wikipedia in a structured manner. It includes a MediaWiki markup parser for an in-depth analysis of page contents. JWPL works with a database in the background, the content of the database comes from a dump, i.e. a static snapshot of a Wikipedia version. JWPL offers methods to access and process properties like in- and outlinks, templates, categories, page text —parsed and plain— and other features. The *Data Machine* is responsible for generating the JWPL database from raw dumps. Depending on what data are needed, different dumps can be used, either including or excluding the Talk page namespace.

The Wikipedia Revision Toolkit (WRT) [Ferschke et al. \(2011\)](#) expand JWPL with the ability to access Wikipedia's revision history. To this end, it is divided into two tools, the *TimeMachine* and the *RevisionMachine*. The *TimeMachine* is capable of restoring any past state of the encyclopedia, including a user-defined interval of past versions of the pages. The *RevisionMachine* provides access to the entire revision history of all Wikipedia articles. It stores revisions in a compressed form, keeping only differences between adjacent revisions. The Revision Toolkit additionally provides an API for accessing Wikipedia revisions along with the metadata like the comment, timestamp and information about the user who made the revision. A more detailed description is provided in appendix [A.1](#)

Wikipedia Miner [Milne and Witten \(2009\)](#) offer a Java-based toolkit to access and process different types of information contained in Wikipedia articles. Similar to JWPL, it has an API for structured access to basic information of an article. Categories, links, redirects and

the article text, plain or as MediaWiki markup, can also be accessed as Java classes. It runs a preprocessed Java Berkeley database in the background to store the information contained in Wikipedia. Wikipedia Miner has a focus on concepts and semantic relations within Wikipedia. It is able to detect and sense-disambiguate Wikipedia topics in documents, i.e. it can be used to wikify plain text. Furthermore, the framework compares terms and concepts in Wikipedia, calculating their semantic relatedness or related concepts based on structural article properties (e.g. in-links) or machine learning. In contrast to JWPL, it cannot be used to access and process the revision history of an article. The capability of its parser is limited, e.g. no templates or infoboxes can be processed.

WikiXRay is a collection of Python and GNU R scripts for the quantitative analysis of Wikipedia data (Ortega, 2009). It parses plain Wikimedia dumps and imports the extracted data into a database. This database is used to provide general quantitative statistics about editors, pages and revisions.

WikiHadoop is a stream-based input format for *Hadoop*⁶⁰, an open-source software for distributed computing. While *WikiHadoop*⁶¹ does not provide any direct support for accessing and processing Wikipedia data, it manages the segmentation of the large data dump for distribution on a compute cluster. Provided with a compressed version of the XML data dump, WikiHadoop decompresses the data on the fly, splits the stream into chunks of single page revisions or pairs of adjacent revisions and feeds these chunks to individual processing units (i.e. mappers). This alleviates the development of code for large scale analysis of Wikipedia data which is only concerned with local information, i.e. the content of a single revision or a pair of revisions.

3.7 Other Wikimedia Projects

While this thesis in general and this chapter in particular focuses on Wikipedia, it is nevertheless important to mention some of its sister projects that closely interact with Wikipedia and share many of its foundational concepts. A full overview can be found on the website of the Wikimedia Foundation⁶².

Wiktionary is a multilingual, collaboratively created, online dictionary that emerged from the desire to exclude linguistic and lexicographic information from Wikipedia articles (Meyer, 2013). Established in December 2002 as a companion to Wikipedia, Wiktionary has grown into a large project of its own right.

⁶⁰<http://hadoop.apache.org>

⁶¹<https://github.com/whym/wikihadoop>

⁶²http://wikimediafoundation.org/wiki/Our_projects

Instead of encyclopedic knowledge, which is the center of attention in Wikipedia, Wiktionary is primarily concerned with word definitions, including additional information such as etymology, pronunciation, and lexical-semantic relations. Going beyond a traditional dictionary, Wiktionary further includes supplemental content such as a thesaurus, a rhyme guide, phrase books or language statistics.⁶³

Like Wikipedia, Wiktionary is available in different language versions, whereas each language version might also describe words from other languages using the native language of the respective version as a point of reference. For example, the English Wiktionary “aims to describe all words of all languages using definitions and descriptions in English”⁶³. In short, each Wiktionary language version is a monolingual dictionary for multiple languages.

Since Wiktionary emerged from Wikipedia, it shares many of its basic principles and technical details, such as user discussion pages for coordination of collaborative efforts and cleanup templates for identifying quality problems. This is why the techniques for quality assessment described in this thesis are, at heart, also suitable for Wiktionary. However, as an online dictionary, Wiktionary has different aims and different quality standards than Wikipedia. Therefore, the underlying quality model which we define for Wikipedia in this work (see chapter 4) has to be adapted to Wiktionary.

Wikidata is a multilingual, collaboratively created, structured knowledge base intended to serve as a central information repository for all Wikimedia projects.⁶⁴ At its core, Wikidata is intended to centralize the management and storage of interwiki links, infoboxes and lists in Wikipedia. Instead of managing the cross-language links between language versions in every article separately, Wikidata provides a central repository of concepts which contain links to all corresponding articles in every language version. This means that a newly added article for a given topic will automatically receive all correct interwiki links when connecting the article to the Wikidata concept. In addition to centralizing interwiki links, Wikidata also contains structured information for many of the concepts stored in the database. A Wikidata-entry for a person, for instance, will typically hold information about their date and place of birth, occupation and other relevant information. This information can be used to automatically fill the infoboxes in all language versions of Wikipedia thus eradicating the problem of out-of-sync information and helping to keep Wikipedia as a whole up to date.

Once deployed in full to all language editions of Wikipedia and other major projects, such as Wiktionary, Wikidata has the potential to increase the consistency and currency of each individual resource and improve the interconnectedness of all Wikimedia projects.

⁶³<http://en.wiktionary.org/w/index.php?oldid=23894302>

⁶⁴<http://www.wikidata.org/w/index.php?oldid=73888516>

3.8 Chapter Summary

In this chapter, we introduced Wikipedia, its main structures and properties, its community and ways to process the large amounts of data it contains. We established that the policies governing Wikipedia and shaping its content are collaboratively defined and change over time. While large parts of these policies are shared across the different language versions, each edition has an individual take on the wiki philosophy which leads to a different culture in each Wikipedia.

Even though Wikipedia contains an almost incomprehensibly large set of rules and guidelines, the basic principles can be boiled down to the five pillars of Wikipedia which build the foundation for a soft security system. A unique characteristic of Wikipedia is the revision history that is kept for every page and which allows keeping track of every change ever made to the encyclopedia. At the same time, the revision history is the reason for the large amount of data Wikipedia sums up to, which makes it difficult to process as a whole.

User communication is mainly performed on the different Talk pages, an unstructured discussion space in dedicated namespaces. Article Talk pages are used to coordinate the article development and discuss the future of an article. User talk pages, on the other hand, are used as the main means of communication between the users. There are different ways to access Wikipedia ranging from direct access to the live databases via a web API over manual processing of downloadable XML dumps to dedicated, database-driven programming interfaces. The best solution depends on the applications' need for data currency and speed.

While the main reason for Wikipedia's success is its policy that everyone can contribute, the same policy also constitutes the greatest challenge. In order to establish Wikipedia as a trustworthy and comprehensive reference work with a quality level equal to edited encyclopedias, Wikipedia needs a quality management process that can cope with the almost anarchic culture that Wikipedia is based on. Taking into account the unprecedented size of the larger Wikipedia editions, a satisfactory solution can only be reached with computational assistance. The remainder of this thesis will therefore sketch how computational methods can assist the community in managing and improving the quality of Wikipedia.

CHAPTER 4

Information Quality

“Quality is never an accident; it is always the result of high intention, sincere effort, intelligent direction and skillful execution; it represents the wise choice of many alternatives.”

— William A. Foster

In this chapter, we discuss the concept of information quality along with theoretical and practical considerations of its measurement. We start with a literature review and identify existing theories and models targeted at describing and quantifying information quality (section 4.1). We then narrow our focus on writing quality and discuss how a model for textually represented information can be derived from a generic information quality model and furthermore examine the factors that pertain to information quality management (section 4.2). We finally review the mechanisms and policies regarding quality assessment and assurance in Wikipedia (section 4.3) from which we derive an article quality model (section 4.4). We conclude the chapter with a summary of our findings (section 4.5).

4.1 Information Quality

Claude Shannon was among the first to develop a quantitative definition of *information* in order to provide a sound theoretical foundation for his model of communication. His definition employs the mathematical uncertainty-measure of *entropy* as a means to quantify the amount of information in a message, i.e. the information content of a message (Shannon, 1948).

While his quantitative definition is useful for the examination of data transmissions and machine-to-machine communication, it falls short of capturing the semantic aspects that we usually associate with the term *information* and which Shannon regards as “irrelevant

to the engineering problem” (Shannon, 1948, p. 379). In fact, Shannon’s definition has no connection to the semantic content of a message at all and even declares that there is potentially more information, i.e. entropy, in chaos and randomness than in structure (Sveiby, 1996). In order to approach a more qualitative definition of the concept of information that more closely resembles the intuitive meaning of the word, one has to take semantic, pragmatic and even aesthetic factors into consideration, which are not as easily captured in codes and numbers.

We will refrain from attempting to define a generic concept of information and rather approximate its meaning indirectly by discussing *information quality*. It is not surprising that quality considerations of something as intangible as *information* are subject to a wide range of uncertainties arising from the different interpretations of the concept, especially given that *quality* itself is also not an easy concept to define.

Information quality (IQ) is generally regarded as a multi-faceted, multi-dimensional concept. In a review of information quality literature, Eppler and Wittig (2000) identify seven basic definitions of information quality, which are often combined in different ways. In summary, high quality information must be *fit for use by information consumers in a particular context*, *meet a set of predefined specifications or requirements*, and *meet or exceed user expectations*. Thus, high quality information provides a particular *high value to the end user*.

Models, Frameworks and Standards. An *information quality model* is a concise system of evaluable criteria which instantiate the aforementioned definitions in a way that they can be incorporated in an *information quality framework*. Such frameworks ideally ground the quality model to an underlying theory, define a scheme for analyzing and solving quality problems and provide metrics for quality measurement (Eppler and Wittig, 2000). *Quality standards* furthermore provide a frame of reference necessary for interpreting the output of the quality measures. In other words, while the *quality model* defines the dimensions along which we measure quality, the *metrics* define how we measure the quality along each dimension and the *standards* define how we interpret the output of these measurements.

As surveys of IQ frameworks (Eppler and Wittig, 2000; Knight and Burn, 2005) show, there is a large dimensional overlap between most available IQ models. These overlaps have often been used as indicators for the most salient and most important dimensions, which, however, ignores the fact that the various models operate at different levels of granularity, in different application contexts and for different types of information.

Since the concept of quality is inherently context-specific, there is no universal IQ model that truly captures all aspects of information quality. However, we can categorize IQ frameworks and models with respect to the extent they have been adapted to particular contexts. We regard a quality framework to be *generic*, if it has not clearly been designed for a specific application context or for a particular information type. We furthermore distinguish

between three adaptation processes, which can be used to customize a generic framework for a given task.

Medium Adaptation: *Adaptation with respect to the representation of the information (e.g. text, video, numerical data) and the way of its distribution or storage (e.g. web, print).*

Application Context Adaptation: *Adaptation with respect to the intended application context in which the information is used or in the context of which it is evaluated. For example, the information quality requirements of a database for medical records are different from the requirements of a public encyclopedia.*

User Adaptation: *Adaptation with respect to the users that interact with the information. This includes factors such as the number of users, their expertise and their way of interacting with the information (e.g. production, consumption or processing of information) (Lee et al., 2002).*

These adaptation processes are not mutually exclusive. In practice, information quality models often exhibit mixtures of adaptations e.g. with respect to a certain medium and application context.

While generic models aim to represent the universal aspects of quality and minimize the adaptation to a specific context or information type by means of more coarse-grained dimensions, specific models employ a more fine-grained set of dimensions to reflect the particular needs of the task at hand. Within a single model, be it generic or specific, all dimensions must be disjunct without any semantic overlap (Rohweder et al., 2008).

Even though generic models are designed without a specific application context in mind so that they can be applied to many different settings, they can only be interpreted when they are contextualized in an application setting.

Wang and Strong Model. One of the most cited models of information quality has been developed by Wang and Strong (1996). Not only has this model been widely used in the last 17 years, it is also recognized as one of the few generic approaches to information quality that take a middle ground between a solid theoretical foundation and practical applicability (Eppler and Wittig, 2000). In a two-stage survey, Wang and Strong asked data consumers with varying backgrounds to identify the individual aspects of data quality along with their perceived importance on a scale from 1 to 9. Overall, they identified 118 different attributes with a high average importance score and a sufficient stability across all participants. Grouping these attributes into higher level categories and merging similar concepts, finally lead to the information quality model shown in figure 4.1. The model consists of 15 quality dimensions that are organized into four categories, the *intrinsic*, *contextual*, *representational* and *accessibility* category.

The intrinsic category captures properties innate to information entities and suggests that data has a certain quality on its own right – independent from its usage, the user or the

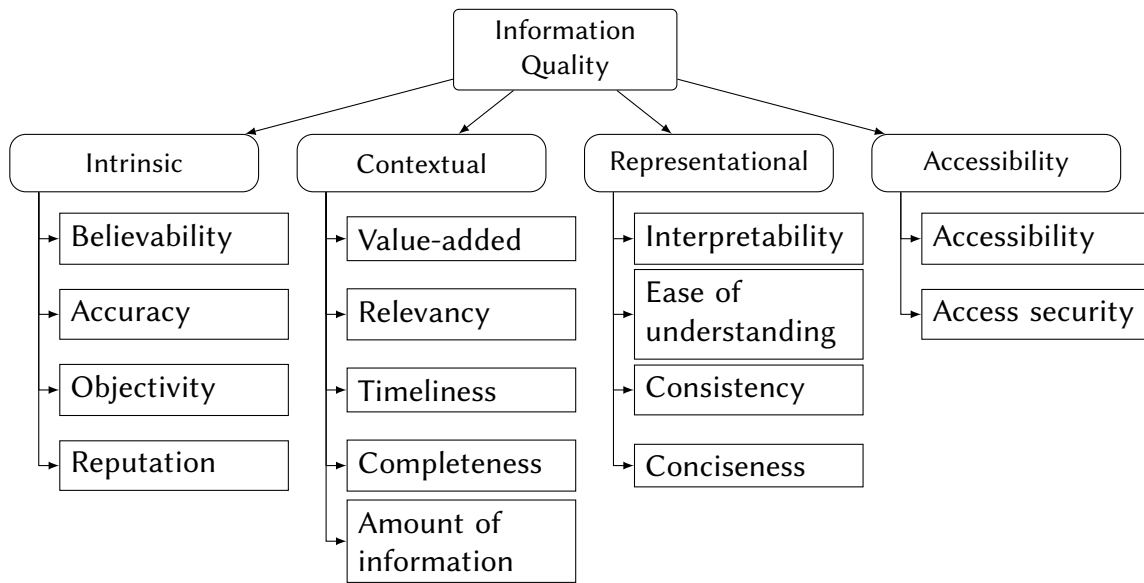


Figure 4.1: Hierarchical information quality model after Wang and Strong (1996)

creator. The contextual category captures the quality in context of the task at hand within which the information entity is used. The representational category furthermore takes into account how the information is represented and whether it can be efficiently processed by the user. Finally, the accessibility category is concerned with the trade-off between security and accessibility, i.e. the ease of accessing information by permitted users.

Due to its strong empirical foundation and its balance between generality and practical applicability, the Wang and Strong model has served as the basis for many IQ frameworks. Among others, the German Association for Information and Data Quality (DGIQ) directly adapted this framework with minor adjustments⁶⁵ as the standard for a user-centric IQ model (Rohweder et al., 2008).

Relationships Between Quality Dimensions. While quality dimensions are supposed to be disjunct within a single model and should be observable in isolation, they are seldom truly independent from each other. Eppler and Wittig (2000) list typical trade-offs between certain pairs of dimensions, such as security vs. accessibility, currency vs. accuracy, or conciseness vs. scope, which suggest that a piece of information, or by extension, a whole information system, can only be optimized for one of the dimensions in each pair. That is, when a document gets more elaborate it becomes less concise. While some applications call for an optimization towards one end of the trade-off spectrum (e.g. security over accessibility), other applications rather seek a balanced calibration.

⁶⁵The dimension *access security* was replaced by *ease of manipulation*, which was originally included in an earlier stage of the Wang and Strong model under the term *ease of operation*. This adaptation stresses the user-focus of the DGIQ-model (Rohweder et al., 2008)

Schaal et al. (2012) give a systematic overview of the relationships between a set of quality dimensions for the social web context. Their model not only contains the trade-off relationship, but also *enabling relationships* such as “*verifiability helps believability*”. However, one can argue that many of the dimensions in an enabling relationship are merely fine-grained distinctions of the same higher-level dimension and are thus positively correlated. That is, verifiability and believability might be considered as the single dimension *trustworthiness*.

The relationships between quality dimensions, particularly the trade-off relations, are important to consider when defining quality standards for the purpose of quality assurance. It is not enough to define what constitutes high quality within every dimension of a model, but it also has to be considered how to balance the quality across dimensions.

Measurability of Quality Dimensions. While quality models build the formal foundation of a quality framework and identify the dimensions along which the quality of information is to be evaluated in a given context, not all of these dimensions are equally well observable in the data let alone measurable automatically. The usefulness of an information quality framework therefore largely relies on the provided metrics for measuring quality along the dimensions the framework defines. The problem of measurability decomposes into four separate aspects (cf. Bizer, 2007, pp.36–39):

Consistency: *How well and how consistently can humans rate quality along each dimension? In other words, is each dimension well enough defined in order to reach reliable judgments with sufficient agreement?*

Subjectivity: *Are quality judgments along a given dimension inherently subject to subjective preference or can it be objectively rated? This aspect will also influence consistency that was mentioned above.*

Operationalizability: *How can we operationalize the judgment along each dimension and what are the indicators on which the judgment is based. In other words, what are the features on which a quality assessment metric can be based for each dimension?*

Interpretability: *Is it possible to map the output of given metric to a quality standard in order to be able to interpret the quality ratings on a scale?*

In the context of automatic quality assessment, operationalizability is the most imminent issue. Yaari et al. (2011) distinguish between measurable and non-measurable criteria, i.e. whether criteria can be reliably assigned by a computer from the text alone without any human intervention, and propose a set of metrics for automatically rating the articles according to the criteria of the former category. Similarly, Stvilia et al. (2007) provide a list of measures that correlate with user judgments from the subset of measurable dimensions in their quality model (see section 4.4). Non-measurable dimensions cannot be operational-

ized directly but need human intervention in the form of manual quality judgments from which correlated aspects can be learned with statistical methods.

Arazy and Kopak (2011) studied the consistency of quality judgment by 270 undergraduate students rating 100 Wikipedia articles on a Likert scale from 1 to 7 along four dimensions⁶⁶ and report poor intra-class agreement levels which did not exceed 0.17. One of the main conclusions drawn from this study is that users have a hard time judging quality within the constraints of an abstract model on a fixed scale. This is not so much a problem of the users not being able to consistently agree on the quality of an article (or in general of an information entity) but more a problem of expressing the subjective quality perception in an abstract rating system. We therefore argue that quality judgments provided in natural language but analyzed according to a well defined model will provide better insights into information quality than having a large crowd of non-experts provide ratings in an abstract form. One of the approaches to assessing article quality presented in this thesis therefore analyzes the discussions of Wikipedia users with respect to information quality judgments.

Information Quality Management. Having defined the terms IQ framework, IQ model and IQ standard earlier in this section, we close with a definition of the processes in which they are employed. *IQ assessment* is the general task of judging the quality of an information entity. Along the lines of the argumentation in this thesis, this is achieved within an IQ framework. *IQ assurance* aims at maintaining a high standard of information quality by continuously monitoring the information quality of all information entities in a resource. This is particularly achieved by identifying and avoiding quality problems. *Quality problems* are defined as violations of a quality standard in any of the quality dimensions defined by the quality model. *IQ improvement* furthermore steers towards a higher quality level, e.g. by providing feedback to the community about quality problems and inconsistencies along with guidelines how to resolve these issues. *IQ management* finally combines IQ assurance and IQ improvement into an integrated process that is tailored towards a particular application context and community. However, it is not only directed at the information entities alone but also at the policies, guidelines and tools responsible for maintaining, securing and storing them (Stvilia et al., 2008) and furthermore shapes the decision making processes involved (Ge, 2009). It has to be noted that these terms are not used consistently across the literature and are often employed interchangeably in different contexts (Eppler, 2003).

⁶⁶A subset of the Wang and Strong model: accuracy, completeness, objectivity, representation

4.2 Text and Writing Quality

A large amount of information that people interact with on a daily basis is represented in the form of text. Since texts are carriers of information, *text quality* can be analyzed in terms of information quality. However, as natural language is very powerful, expressive and subject to many subtleties, applying a generic, coarse-grained information quality model to text will leave many facets unexplained and will thus be insufficient.

A quality model adapted for textual information should therefore consider the peculiarities of the medium. Above all, this involves the incorporation of a notion of well-writtenness, or, in other words, the *writing quality* of a text. While text quality models capture the overall quality of textually represented information, writing quality is a subordinate concept that mainly reflects the quality of its representation. Thereby, it captures both formal aspects of language correctness, sometimes also referred to as *linguistic quality*, and creative aspects of language use, such as the development of ideas within a text or the effective use of rhetorical devices. In the following, we discuss the major aspects of writing quality which we will further break down in section 4.4 when introducing our Wikipedia article quality model.

Language correctness. In its essence, language correctness concerns the proper use of the lexicon and the compliance of a text with standards and conventions prescribed by the grammar of the given language. However, since languages are not static systems and rather subject to constant development, correctness can only be partially derived from a prescriptive grammar and lexicon and has to be evaluated in the light of the dynamics of language use. Furthermore, the binary notion of correctness is insufficient for real world texts and should rather be regarded as a graded scale of language acceptability (Gordesch and Dretzke, 1998). For example, a spell checker might define the colloquial expression *wassup*⁶⁷ as incorrect, although the term might very well be acceptable in the context of social media. Also, non-standard grammatical and syntactic constructions might be considered incorrect in one context while being acceptable in another. Therefore, automatic spell and grammar checkers are helpful resources for text quality assessment, especially in the context of encyclopedias, which aim for standard language usage. However, since collaboratively created texts will always reflect the dynamics of language, the concept of *language correctness* has to be taken with a grain of salt.

Writing Traits and Rubrics. Beyond mere correctness, academic standards in the language arts define *traits* of writing quality with the goal of standardizing the assessment of student writing and giving students feedback on their state of writing proficiency. A

⁶⁷Meaning *what's up*, a simplified form of greeting, see <http://www.urbandictionary.com/define.php?term=Wassup>, accessed on April 4th, 2014

widely adapted framework is the six traits scoring rubric for writing assessment described by Spandel (2012). Derived from a large scale analysis of student essays in order to find common characteristics of good writing, the six traits aim to deliver a guide for assessing and teaching writing at any level of proficiency and for every audience and text type. In short, the six traits capture the following aspects of a text:

Ideas and development: *Development of ideas, clarity and focus of the text, level of detail.*

Organization: *Order, presentation and structure of the text.*

Voice: *Choice of an appropriate⁶⁸ tone and stylistic level.*

Word choice: *Choice of appropriate⁶⁸ words and register.*

Sentence fluency: *Rhythm and flow of language; readability and understandability.*

Conventions: *Language correctness.*

While they have been developed for language teaching, the six traits have successfully been applied in the area of language technology as a base model for automatically assessing the quality of scientific journalism (Louis, 2013). Following the argumentation of Louis, we argue that these rubrics can be good indicators of writing quality in the context of an encyclopedia if they are sufficiently adapted to the genre. In our Wikipedia article quality model described in section 4.4, many of the dimensions in the writing quality category are therefore based on the theoretical foundation of the six traits.

Readability. From a computational perspective, *readability* is among the oldest and best researched aspects of writing quality. In order to determine the level of reading competency needed to understand a text and to quantify the clarity of writing, many of the early readability metrics rely on textual surface features. To this end, shallow properties, such as the average number of words per sentence or characters per word, are combined in different formulas (Kincaid et al., 1975; Smith and Senter, 1967; Coleman and Liau, 1975; Flesch, 1948; McLaughlin, 1969; Gunning, 1969). In a recent empirical study (Pitler and Nenkova, 2008), these shallow features have not proven to correlate highly with human readability judgments. They could rather show that lexical, syntactic, semantic or discourse features are more predictive of how well a text is written with respect to reading ease, since these aspects include text organization and lexical difficulty in the equation. Ultimately, readability assessment is an audience specific endeavor, since it aims to determine the comprehensibility of a text for a particular target audience. This notion, however, is incompatible with writing quality assessment intended for a general audience. We therefore have to reinterpret readability scores as an absolute measure of complexity and learn from the data which readability level is regarded as appropriate by the majority of the readers (Louis, 2013).

⁶⁸Appropriate for the genre of the text, the audience, the publication medium and also the writing proficiency of the author as indicated in the six traits guidelines.

Text Organization. While readability is mainly concerned with the sentence level, text organization on the document level is responsible for how well a text is readable as a whole and how easy the argumentation can be followed.

Coherence describes the internal consistency of a text exhibited by a “continuity of senses” (de Beaugrande and Dressler, 1981, p. 84). This means that concepts and arguments must be logically connected in order for the recipient to be able to make sense of the text as a whole. On the surface level, the cohesion of a text captures how well the sentences in a text hold together. Cohesion helps to follow the argumentation in a text and is mainly reached with the help of coreference chains, ellipsis, substitution, conjunction and lexical chains (Halliday and Hasan, 1976). While a cohesive text has a higher probability of being coherent, it is still possible for a text to be both cohesive and incoherent.

Motivated by theories of discourse structure (Grosz and Sidner, 1986) and local coherence between adjacent sentence pairs (Grosz et al., 1995), many metrics have been developed that incorporate discourse connectives and coreference analyses into measures of text organization. Louis (2013) gives a detailed overview of related work in this area.

Moreover, psycholinguistic coherence measures, such as *Coh-Matrix* (Graesser et al., 2004, 2011; McCarthy et al., 2006), aim to reflect the coherence of texts with a combination of textual features and metrics and thereby seek to replace the readability metrics on the sentence level while linking the output to psycholinguistic theories. *Coh-Matrix*, as one example, combines 54 metrics ranging from shallow surface features similar to the ones employed in readability assessment over latent semantic analysis and frequency based lexical models to a cue-based analysis of discourse connectives.

While most of the above mentioned approaches rely on a strict linguistic theory of discourse structure, the statistical revolution in NLP also produced new data-driven approaches that aim to determine patterns in the discourse structure from the data. Barzilay and Lapata (2005), for example, apply the centering theory to automatically learn entity transitions between adjacent sentences and thus overcome the need for explicit computation. Louis (2013) furthermore proposes a data-driven approach for measuring the intentional structure of writing by using syntax as a rough proxy rather than explicitly annotating the intentional structure for each text genre manually.

4.3 Quality Management Mechanisms in Wikipedia

In an open, collaborative environment such as Wikipedia, no central institution or committee regulates how quality is to be measured and what standards are to be used as a frame of reference. The quality management process is rather defined by an agglomeration of constantly changing guidelines and policies which are largely fragmented and distributed over different places in Wikipedia.

Three central directories give a comprehensive overview of the available guidelines⁶⁹, policies⁷⁰ and best practices⁷¹.

Distinguished Content. The central frame of reference regarding content quality in Wikipedia is the *distinguished content certification*. The highest level of distinction in the English Wikipedia is the *featured content level*, which means that the content meets all quality criteria for Wikipedia content⁷², has been evaluated in peer review, and is thus eligible to be *featured* on the Wikipedia main page on a rolling basis. This certificate not only exists for articles, but also for other kinds of Wikipedia content, such as lists, pictures, sounds, portals or topics⁷³. The *featured article criteria*⁷⁴ state that an eligible article, in addition to abiding to the aforementioned content policies, should meet the following requirements:

- Be well written, comprehensive, well researched, neutral and stable
- Have a concise lead section
- Be well structured, well referenced and well illustrated
- Have an adequate length and an appropriate level of detail

While most of these criteria are illustrated by accompanying guidelines, there is no exact definition of what differentiates, say, very good writing from mediocre writing and where one should draw the line. Rather than employing a fixed, external frame of reference for quality assessment, the quality judgment is done by comparison to other articles with featured status. Since these existing featured articles are further improved over time, the *standard-by-comparison* rises, making it more and more difficult for new articles to qualify for featured status. This becomes most evident when comparing the featured articles from years ago with today's featured articles. If a featured article gets demoted, i.e. the featured status is removed in another peer review process, this is rarely caused by a degradation of its quality. The article rather has not improved as fast as the collective standard has risen. Due to this fact, and due to the complexity of the peer review process described below, only a very small number of articles has featured status (4,782 or about 0.1% in the English Wikipedia).

Articles that largely comply with the featured article criteria but fall short in some of the categories can qualify for *good article status*, a lower level distinction that only exists for articles. As of the time of writing, less than 0.3% of all articles in the English Wikipedia have this status.

⁶⁹<http://en.wikipedia.org/wiki/WP:LGL>

⁷⁰<http://en.wikipedia.org/wiki/WP:LOP>

⁷¹<http://en.wikipedia.org/wiki/WP:MOS>

⁷²<http://en.wikipedia.org/wiki/WP:CONPOL>

⁷³<http://en.wikipedia.org/wiki/WP:FC>

⁷⁴<http://en.wikipedia.org/wiki/WP:FACR>

| Quality | Importance | | | | | Total |
|--------------------|------------|---------|---------|-----------|------------|-----------|
| | Top | High | Mid | Low | Unassessed | |
| A | 178 | 316 | 512 | 281 | 70 | 1,357 |
| B | 10,399 | 19,847 | 29,877 | 21,626 | 12,421 | 94,170 |
| C | 7,868 | 22,263 | 49,194 | 58,698 | 32,800 | 170,823 |
| Start | 14,978 | 63,960 | 255,432 | 565,328 | 224,062 | 1,123,760 |
| Stub | 3,894 | 26,742 | 188,795 | 1,337,343 | 870,398 | 2,427,172 |
| List | 2,325 | 8,732 | 24,082 | 61,827 | 49,701 | 146,667 |
| Featured Articles | 1,002 | 1,514 | 1,401 | 794 | 162 | 4,873 |
| Featured Lists | 133 | 511 | 593 | 542 | 125 | 1,904 |
| Good Articles | 1,637 | 3,736 | 7,303 | 6,867 | 1,509 | 21,052 |
| Overall Assessed | 42,414 | 147,621 | 557,189 | 2,053,306 | 1,191,248 | 3,991,778 |
| Overall Unassessed | 117 | 319 | 1,379 | 16,001 | 466,465 | 484,281 |
| Total | 42,531 | 147,940 | 558,568 | 2,069,307 | 1,657,713 | 4,476,059 |

Table 4.1: Number of articles per WikiProject quality level and importance category in the English Wikipedia according to <http://tools.wmflabs.org/enwp10/cgi-bin/table2.fcgi> accessed on 04 April 2014. A,B,C, Start, Stub, List = WikiProject Quality Grades

Considering that less than 0.4% of Wikipedia articles received a distinction for excellent content, it is safe to assume that they cannot be representative of all the good content in Wikipedia.

Peer Review. Featured and good articles are determined in a peer review process⁷⁵. An article first has to be nominated by a community member and will only be accepted if no major cleanup templates are assigned to the page. If accepted, the article is listed among the featured or good article candidates. In a next step, reviewers are recruited from within the community to manage the further process. Reviewing takes place on a dedicated sub page in the Talk namespace of the article. In collaboration with a group of volunteers, open issues are addressed until a final verdict is reached. In case the article is eligible for featured or good status, the corresponding category is assigned. Otherwise, the case is closed and peer review can be restarted after two weeks time. Peer reviews can also be requested in other contexts related to quality assurance and assessment.

WikiProject Article Quality Ratings. Since the peer review process described above is too complex and labor intensive for judging the quality of every article in Wikipedia, the problem task has been distributed across the individual WikiProject subgroups (see chapter 3.3.4). Assuming that WikiProject members are experts in their subject area, they are

⁷⁵<http://en.wikipedia.org/wiki/WP:REVIEW>

asked to rate the importance of an article for their field along with its quality on a predefined scale. Article importance ratings range from *top* to *low* importance, while the quality levels range from A to C in descending order. Overly short or newly created articles can furthermore receive a *stub* or *start* grade respectively. Table 4.1 shows the numbers of articles per quality level and importance category. It additionally lists the number of pages that received a particular distinguished content certification (see above).

The quality judgments are centrally gathered by a bot (WP 1.0 bot) and used to monitor the overall quality status of Wikipedia. Furthermore, the ratings are the basis for compiling Wikipedia offline releases, which are created on an irregular basis.⁷⁶ While the WikiProject quality assessment scale is fixed in terms of quality levels, no exhaustive criteria are defined for each grade⁷⁷. Moreover, while featured and good article status is assigned in a well-defined peer review process, the WikiProject quality assessment is subject to local customs of the respective WikiProject. Consequently, articles with the same quality level assessed by members of different WikiProjects might considerably differ in their actual quality. In short, the assigned quality levels are not necessarily comparable across Wikipedia. Finally, even though most articles in the English Wikipedia have been rated with this rating scheme at least once, there is no information available as to how recent a certain rating is. It largely depends on the activity of the members of a WikiProject how well the ratings reflect the current state of affairs.

User Feedback. While distinguished content and WikiProject quality ratings are grades assigned by active Wikipedia authors, a large fraction of Wikipedia users are purely passive, i.e. they exclusively read but do not contribute to the encyclopedia. In order to incorporate the opinions and views of these users into the quality management process, the *Article Feedback Tool* (AFT) has been developed. It allows the whole Wikipedia community to evaluate articles along the dimensions *Trustworthy*, *Objective*, *Well written* and *Complete* on a five-star scale. In addition to the actual article ratings, the users were furthermore encouraged to provide information about their knowledge in the subject area of the article in order to put these anonymous ratings into perspective. The user interface is displayed in figure 4.2. Even though the ratings ask for judgments on a five-star scale, users tend to rate with extreme scores that merely reflect a binary classification in each dimension (good vs. bad) (Flekova et al., 2014). Additionally, the high correlation between the four dimensions renders separate analyses difficult.

Cleanup Templates. As described in chapter 3.2.4, the template system in Wikipedia is not only used to embed recurring content into Wikipedia pages, but also as a tagging system

⁷⁶<http://en.wikipedia.org/wiki/WP:1.0>

⁷⁷ <http://en.wikipedia.org/wiki/WP:ASSESS> gives a brief description of each category along with a sample article

Figure 4.2: The rating interface of the Article Feedback Tool (v4) as displayed on the bottom of an article.

for various applications. One of these applications is marking open issues in articles that need attention of a contributor. While these cleanup templates leave a note on the tagged page, they also enlist that page in the corresponding cleanup category. Furthermore, short notes can be added as parameters to the template in order to document the issue and provide additional details. Thus, the cleanup template system constitutes an issue tracker for quality assurance and is used as the basis for our approach to automatically identify quality flaws in Wikipedia articles that is described in chapter 5.

Flagged Revisions and Pending Changes. As described in chapter 3.4, the flagged revisions extension and, in the English Wikipedia, the pending changes extension are used to hide new changes from the general public until they have been reviewed by trusted community members. While this review process is, in practice, not really suitable to improve the article quality, it helps to prevent vandalism and the associated deterioration of article quality.

Article Discussions. Even though the Wikipedia Talk pages (see chapter 3.5) are not, by design, instruments for quality management, the article Talk pages are nevertheless the central platform for any communication regarding article development and work coordination. On these pages, both the active contributors and the more passive readers exchange their thoughts on how the article can best be improved and share criticism regarding its quality. However, since these discussions are largely unstructured and often spread over several Talk archives, it is not easy to keep track of decisions made in the past. Furthermore, a study by Schneider et al. (2011) has shown that new users are easily confused by the lack of discourse structure and therefore do not contribute in larger discussions. We therefore

propose an approach in chapter 6 to automatically analyze these discussions in order to provide structured feedback regarding quality management decisions.

Overall, while having been successful in the past to make Wikipedia a reliable information source and one of the central reference works on the Internet, the quality management mechanisms based on community decisions alone are unlikely able to cope with the exploding size of Wikipedia. It is therefore necessary to provide computational assistance without patronizing the users and imposing too many restrictions on the community, their work and their decisions.

4.4 An Article Quality Model for Wikipedia

As we have seen in the outset of this chapter, the concept of information quality largely depends on different contexts, for example the intended application in which an information entity is supposed to be used, the target audience, or the form of representation. Therefore, a quality model for Wikipedia articles has to consider the characteristics and purpose of an encyclopedia, has to reflect its collaborative and intercultural nature and has to address the peculiarities of its representation.

Several attempts have been made to adapt existing information quality models and frameworks to encyclopedias in general and to Wikipedia in particular (Crawford, 2001; Stvilia et al., 2007, 2008; Lichtenstein and Parker, 2009). Also the Wikipedia community formed a *Quality Task Force* (QTF) in order to develop a quality model that can measure “the ability of a [W]ikipedia article to meet the expectations and needs of the article’s target audience, i.e. the readers of the article”⁷⁸. In four categories, the QTF model defines quality dimensions capturing *requirements of content*, *demand*, *form* and *the project*. While the first three dimensions consider quality aspects of individual articles, the latter applies to greater structures, such as whole subject areas and the integration of individual articles within these areas.

While these attempts succeed in incorporating the requirements of a collaboratively created encyclopedia with respect to intrinsic and contextual information quality, the representational aspects fall short in most of these models. The quality of writing, while being a key aspect for a text resource such as Wikipedia, is not clearly distinguished from the intrinsic quality aspects or merely represented in an undifferentiated, aggregate form. As the only exception, the QTF model attempts to explicitly include representational aspects of Wikipedia articles. However, the model has never left the stage of a working definition and it is neither based on a sound theoretical foundation nor integrated in a quality assessment framework.

⁷⁸<http://strategy.wikimedia.org/w/index.php?oldid=65341>

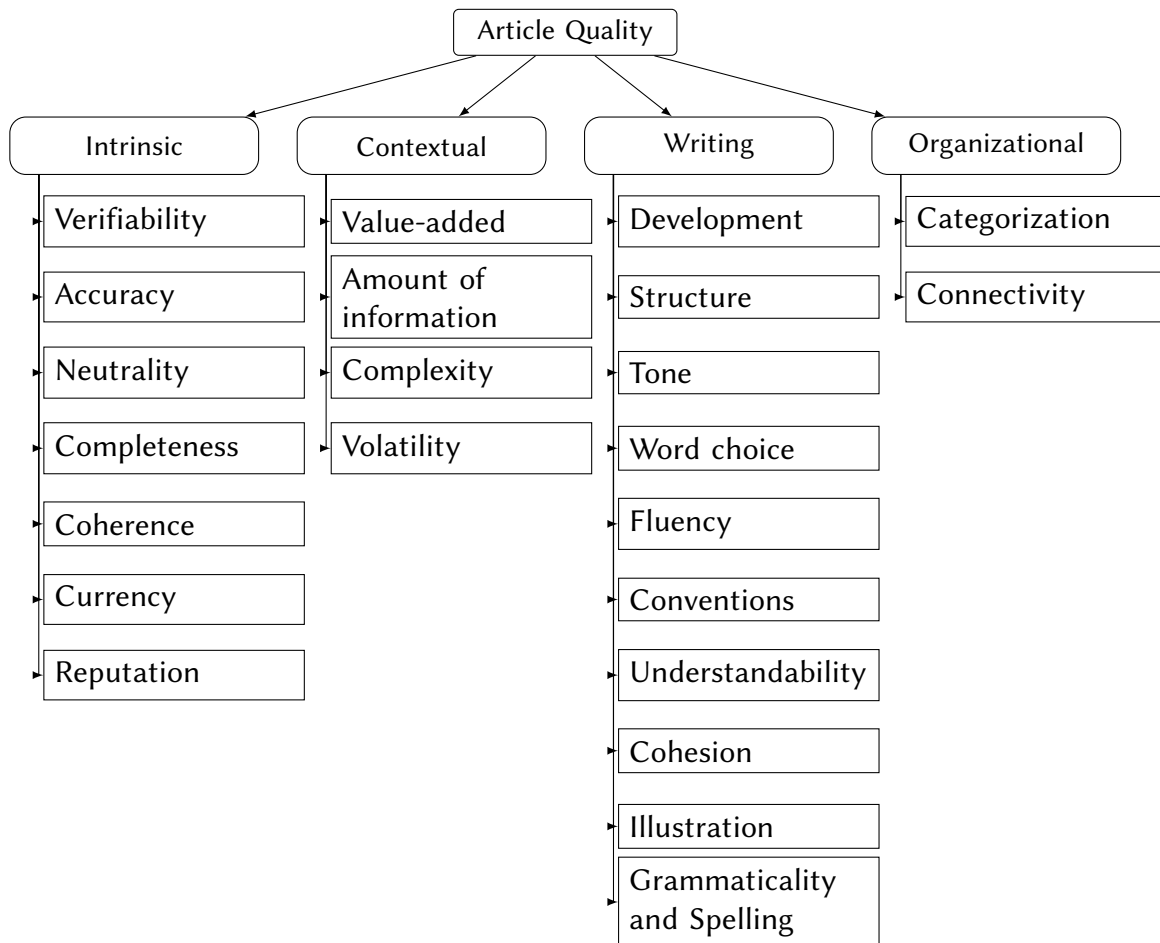


Figure 4.3: Proposed model for article quality in Wikipedia

We therefore build upon the previously discussed generic model of information quality by [Wang and Strong \(1996\)](#) and, under consideration of the related adaptations of this model for Wikipedia, define a unified model of article quality paying special attention to textual properties. This model is supposed to represent the writing quality of Wikipedia articles in the sense that we defined in section 4.2. It is supposed to serve as an orientational map for quality management that can be used to identify aspects of information quality to be monitored and assessed by available methods and mechanisms and also indicate gaps in the coverage of existing quality assurance processes. We will later refer to this model when we introduce our automatic quality assessment methods in the chapters 5 and 6.

Similar to [Wang and Strong](#), we distinguish between four categories of quality dimensions. Following their rationale, we define a category of intrinsic quality and contextual quality. While the former captures the internal characteristics of the information that is expressed by an article, the latter focuses on its appropriateness for the audience, medium and application. Instead of a generic category of representational quality, we account for

the textual nature of Wikipedia with a dedicated writing quality category that includes linguistics and stylistic properties. Finally, while the first three categories assess individual articles in isolation, the fourth organizational category focuses on their integration within the wider confines of Wikipedia. Figure 4.3 shows an overview of the model. In the following, we give a short description of each dimension in the model.

4.4.1 Intrinsic Article Quality

Following Wang and Strong and their category of intrinsic data quality, *intrinsic article quality* captures the internal characteristics of the information contained in an article and is largely detached from its representation or application. Judging article quality along the dimensions in this category demands the greatest level of knowledge about the article topic.

Verifiability: Originally defined as *believability* by Wang and Strong, we define *verifiability* to assess how well the information represented is referenced and can thus be verified. This both affects the authority and quality of any given sources as well as the absolute number of sources contained in the article and the relative coverage of the article content by references.

Accuracy: Refers to the factual correctness and the preciseness of an article.

Neutrality: A neutral article is not supposed to take particular sides and should provide a balanced view on a subject. Issues regarding article neutrality are often discussed under the term NPOV – the neutral point of view. Even though there is a general distinction between neutrality and objectivity, i.e. an article can be objectively written while not being neutral – we subsume both concepts under the same dimension.

Completeness: The dimension *completeness* is strongly related to the *amount of information* on the contextual level. However, in contrast to the amount of information, completeness is not concerned with the verbosity of the article, but rather with how well the article topic is covered. This dimension is also present in the user rating dataset described in section 4.3. However, in this dataset, completeness scores mainly correlate with the article length, which is an important factor but is by no means sufficient to rate the quality along this dimension reliably. Judgments along this dimension demand extensive knowledge of the subject area in order to identify coverage gaps.

Coherence: Describes the internal consistency of the information. Coherent texts exhibit a “continuity of senses” (de Beaugrande and Dressler, 1981, p. 84) meaning that concepts and arguments must be logically connected in order for the recipient to be able to make sense of it as a whole. Coherence is strongly related to cohesion (see below), which captures the linguistic aspects of coherence (Halliday and Hasan, 1976).

Currency: Captures how up to date the information is, i.e. whether the article reflects the current state of affairs and the current state of knowledge.

Reputation: While the reputation of traditional publications is often associated with the respective reputation of the author and the publisher, reputation in the Wikipedia context is more concerned with the trustfulness of the sources from which the information was taken. This dimension is strongly related to verifiability and captures the quality of references rather than the article coverage with references.

4.4.2 Contextual Article Quality

The category of contextual article quality consists of dimensions that capture how well the article fits into an encyclopedia and how well it satisfies the typical requirements of the audience. First and foremost, the article needs to fill a knowledge gap within the encyclopedia and concisely describe its subject on an appropriate level of detail.

Value-added: The added value of an encyclopedic article is mainly that it delivers relevant and concise information that is most likely of interest for the typical reader without repeating information that is already available somewhere else in the encyclopedia. It is therefore also often described with the duality of *informativeness* and *redundancy* (Stvilia et al., 2007, 2008).

Amount of information: Related to the dimension *completeness*, the amount of information relates to an adequate level of verbosity of the article. An article should be as verbose as necessary while being as brief as possible in order to fulfill all requirements along the other quality dimensions.

Complexity: This dimension relates to the level of abstraction and level of detail at which the topic of the article is described. It is strongly related to the dimension of understandability, which covers the linguistic complexity of the article. The present dimension rather captures the level of complexity at which the information is presented.

Volatility: The volatility of an article is determined by the stability of its content. Depending on the article topic, a certain level of constant revision is needed to fulfill the requirements of the *currency* dimension. Apart from that, a high quality article should not be subject to frequent larger revisions.

4.4.3 Article Writing Quality

In accordance with our previous description of writing quality, this category captures how well an article text is developed and subsumes both linguistics and stylistic properties. Similar to previous work on scientific journalism (Louis, 2013), the dimensions in this category are loosely based on the six traits scoring rubric for informative texts (see discussion in

section 4.2) with genre-specific adaptations and expansions for encyclopedic texts under consideration of the Wikipedia Manual of Style.⁷⁹

Development: Ideas expressed in the article have to be logically organized.

Structure: The article should be well structured according to the Wikipedia style guidelines, use sectioning with meaningful headlines and paragraphs within the sections.

Tone: Tone of the text should be suitable for a formal, informative text and neither be too casual and intimate nor too stilted. It should be neutral and distant rather than opinionated and engaging.

Word choice: The right choice of words involves the selection of an appropriate register (in accordance with the *tone* dimension) and should account for precise and natural sounding language. According to the requirements of the *understandability* dimension, the use of technical terms should be limited to the necessary minimum.

Fluency: The article should fluently read as a single text rather than represent a collection of independent text snippets related to the same topic.

Conventions: The article has to maintain the conventions defined in the Wikipedia Manual of Style regarding the aspects such as abbreviations, capitalization and punctuation.

Understandability: Strongly related to the dimensions *complexity* and *word choice*, understandability captures linguistic aspects such as the syntactic complexity of the text. It subsumes concepts of readability and reading ease, which have been discussed in section 4.2

Cohesion: Cohesion refers to the linguistic aspects of the related dimension of *coherence* and captures how well the sentences in a text hold together. Cohesion helps to follow the argumentation in a text and is mainly reached with the help of coreference chains, ellipsis, substitution, conjunction and lexical chains (Halliday and Hasan, 1976).

Illustration: While the adequate illustration of a text with images, tables or graphs in order to support the effective delivery of an article's message is an extratextual aspect, we still regard it as a trait of writing quality, since it creates a direct link between the textual and the visual level. This aspect has recently also been incorporated in the six traits rubric in form of an additional *presentation* category (Spandel, 2012).

Grammaticality and Spelling: This dimension refers to the previously discussed aspect of language correctness and comprises the correct use of grammar and spelling. It furthermore defines how language varieties and the use of non-standard language are to be treated. For instance, the English Wikipedia allows both American English and

⁷⁹<http://en.wikipedia.org/wiki/WP:MOS>

British English to be used but requires each variety to be employed consistently within a single article.

4.4.4 Organizational Article Quality

While the previously described categories of quality dimensions assess individual articles in isolation, the organizational category is concerned with the integration of the articles within the wider confines of Wikipedia.

Categorization: Wikipedia makes use of a comprehensive system of categories (see section 3.2.2) which help to improve the navigation through the encyclopedia, improve findability of the articles and help to automatically create topical lists and overviews. Therefore, the correct and appropriate categorization of articles is vital for the overall quality of Wikipedia.

Connectivity: Wikipedia articles are hypertext documents that strongly rely on their interconnection with other articles. In order to avoid redundancy across Wikipedia while maintaining the understandability of individual articles by providing all necessary information, articles rather link to existing content than reproducing said content. Therefore, when judging the quality of a single article, we have to consider its integration within the wider scope of a superordinate WikiProject, the whole Wikipedia, or even within the network of Wikimedia projects.

4.5 Chapter Summary

In this chapter, we discussed the concept of information quality and its application to information quality management. We have established that information quality, in the broadest sense, is a measure of the “fitness for use” of an information entity in a given application scenario. While it is not possible to define a single universal model of information quality, the models differ in how far they have been adapted to a particular application, medium or user group. The notion of text quality refers to an information quality model for textually represented information which particularly takes the writing quality of a text into account. In order to construct an information quality model for Wikipedia articles, we reviewed the existing mechanism for information quality management in Wikipedia to gain an overview how the concept of quality is interpreted in this community. Based on the widely accepted generic IQ model by Wang and Strong (1996), we then described a hierarchical article quality model with 23 dimensions in four categories that particularly includes writing quality as a major component. The role of this model in the remainder of this thesis is to provide a means of orientation with respect to the aspects of quality that can be assessed with our proposed methods and also show the gaps that remain.

CHAPTER 5

Quality Flaw Detection in Wikipedia Articles

“Conceal a flaw, and the world will imagine the worst.”

— Marcus Valerius Martial

A major part of information quality management is quality assessment, the goal of which is to measure the quality of a given information entity according to a predefined quality model and standard. Often, the output of the quality assessment process is an abstract score that gives no rationale regarding the concrete quality problems of the information entity and therefore cannot directly contribute to improving the information. In this chapter, we discuss our approach to quality flaw detection in Wikipedia, which identifies particular violations of a quality standard and therefore directly alleviates quality improvement efforts.

We first give an overview of our motivation (section 5.1) and proceed with a formal introduction of the concept of quality flaws and how they are manifested in Wikipedia (section 5.2). We then introduce two corpora of Wikipedia articles with selected quality problems and discuss the problem of selecting reliable documents while avoiding a topic bias (section 5.3). Finally, we investigate how these corpora can be used to automatically identify quality flaws in unseen articles (section 5.5) and examine a method to mine flaw corrections from the article revision history (section 5.6). We conclude the chapter with a discussion of the limitations in the predictability of cleanup flaws (section 5.7) and a summary of our findings (section 5.8).

5.1 Motivation and Overview

In the previous chapter, we have introduced the concept of information quality and have described the major aspects of information quality management in the context of a collaboratively created encyclopedia. While quality assessment is an important part of information quality management, it is not enough to just quantify the quality of an information entity – i.e. a Wikipedia article – with an abstract quality score or by assigning coarse grained labels identifying exceptional content. This, however, has been the central approach of related work, which mainly focused on the prediction of *good* and *featured article* labels (Wilkinson and Huberman, 2007; Lipka and Stein, 2010; Javanmardi and Lopes, 2010) or automatically assigning the community defined *WikiProjects article quality grades* (Hu et al., 2007; Rassbach et al., 2007; Hasan Dalip et al., 2009; Han et al., 2011b,a). The main problem with these approaches is that they do not provide any rationale why an article received a particular quality rating, what the quality problems are and how to improve the article and its quality. We argue that it is rather important to identify concrete quality problems in order to inform the Wikipedia community where their efforts are most needed and to assist them in improving the overall quality of the encyclopedia.

In this chapter, we present an approach for identifying quality flaws in Wikipedia articles based on supervised text classification using cleanup templates assigned by Wikipedia users as training data. Quality flaw detection constitutes a data-driven quality management strategy directly aimed at assessing and improving the data. The task has first been introduced by Anderka et al. (2012) who showed the feasibility of this approach for the ten most frequent cleanup templates. In our experiments presented in this chapter, we first extend the scope of the task to a wider set of more subtle quality flaws in the categories *neutrality* and *style*, which have been identified as particularly important for article quality by the Wikipedia community. We put particular emphasis on the data sampling techniques employed, because an analysis of related work has shown that this task is prone to a severe topic bias in the training data. This results in overly optimistic cross-validated classification results that do not realistically reflect the classifier’s true performance. We therefore present a technique to factor out the topic bias and extract reliable training instances from the article revision history. We furthermore show how this approach can be extended to mine quality flaw corrections from the history.

The main contributions of this chapter can be summarized as follows:

Contribution 5.1: *We present a new corpus of neutrality and style flaws mined from the English Wikipedia. The corpus contains both articles with the particular flaws and documents that are reliable examples for articles without these flaws. (section 5.3)*

Contribution 5.2: *We identify that the quality flaw identification task based on cleanup template detection is prone to a topic bias that results in unrealistically high cross-validated evaluation results that do not reflect the classifier’s real performance on*

real world data. We furthermore propose a data sampling approach that is able to avoid this bias in the training data. (section 5.3)

Contribution 5.3: We introduce *FlawFinder* – a system for supervised text classification designed for quality flaw detection (section 5.4). While *FlawFinder* has been developed particularly for the task described in this chapter, it can be applied to general text classification problems and has been adapted as a general purpose text classification framework that is described in appendix A.2

Contribution 5.4: We evaluate the performance of *FlawFinder* trained both on the corpora used by related work and on the newly created neutrality and style corpora and perform a detailed error analysis (section 5.5)

Contribution 5.5: We finally describe an approach for mining a corpus of quality flaw corrections from Wikipedia’s article revision history. (section 5.6)

5.2 Quality Flaws in Wikipedia

While aggregated quality scores in Wikipedia, as they are represented by the featured and good article labels or the WikiProject quality grades (see chapter 4.3), might separate high quality articles from the rest, they fail at representing the intermediate range on the quality scale and do not provide any rationale for the assignment of a particular label, i.e. they do not identify the *quality problems* of an article. Quality problems can be defined as violations of the quality guidelines, i.e. deviations from the quality standard in any quality dimensions.

Cleanup templates (see chapter 3.2.4), on the other hand, represent to-do markers assigned to articles by Wikipedia users in order to identify concrete shortcomings and deficiencies of an article that have to be fixed by the community. These human-assigned labels are therefore very good indicators for quality problems and thus a promising resource for training quality flaw classifiers which can identify these problems in unseen articles. In contrast to labels that simply assign a quality grade to an article, these atomic markers identify single problems which directly give actionable feedback to the community with respect to possible improvements. In turn, the aggregated set of all quality flaw markers assigned to an article can also give an overall impression of its quality status.

In the following, we give an overview of the properties of quality flaws in Wikipedia, formally define the task of quality flaw detection and analyze the reliability of cleanup templates as quality flaw markers.

5.2.1 Properties of Quality Flaws in Wikipedia

The system of cleanup templates, like most organizational structures in Wikipedia, has grown organically and is still subject to constant change. Rather than being a well-drafted taxonomy which corresponds to a quality model as the one which we defined in chapter 4, it

is a loose agglomeration of tags with different granularity and semantic overlaps. Therefore, we distinguish between *quality flaws* on a conceptual level and the *cleanup templates* on a concrete level with the templates being manifestations of the flaw. In the following, we discuss the characteristics of both cleanup templates and flaws and examine how they are related to each other.

5.2.1.1 Template Scope

Since quality flaws are represented by cleanup templates, it is important to consider the scope of each template in order to locate its respective point of reference. We distinguish between three different scopes. *Inline-templates* are placed directly in the text and refer to the sentence or paragraph they are placed in. Templates with a *section* parameter refer to the section they are placed in. The majority of templates, however, refer to a whole page. Figure 5.1 shows examples for each scope.

The consideration of template scope is of particular importance for quality flaw recognition problems. For example, the presence of a cleanup template which marks a single section as *not notable* does not entail that the whole article is not notable. In other cases, however, inline- or section-scope templates can be extended to the whole page. For instance, if a section is marked to contain original research, this also holds true for the complete article. In these cases, templates with a narrower scope help to locate the problems in the tagged article.

5.2.1.2 Template Clusters

Since several cleanup templates might represent different manifestations of the same quality flaw, there is a 1 to n relationship between quality flaws and cleanup templates. For instance, the templates `pov-check`⁸⁰, `pov`⁸¹ and `npov language`⁸² can all be mapped to the same flaw concerning the neutral point of view of an article.

The degree of similarity between two templates can differ. We can roughly distinguish three cases:

- (1) Two cleanup templates are *fully synonymous* if one template redirects to the other.
- (2) Two cleanup templates are *similar*, if they capture the same problem type but differ in scope or granularity/specificity.
- (3) Two cleanup templates are *unrelated*, if they are neither synonymous nor similar.

Fully synonymous templates are easy to determine automatically by extracting the redirects from and to the template information pages. The names of synonymous templates are

⁸⁰The article has been nominated for a neutrality check

⁸¹The neutrality of the article is disputed

⁸²The article contains a non-neutral style of writing

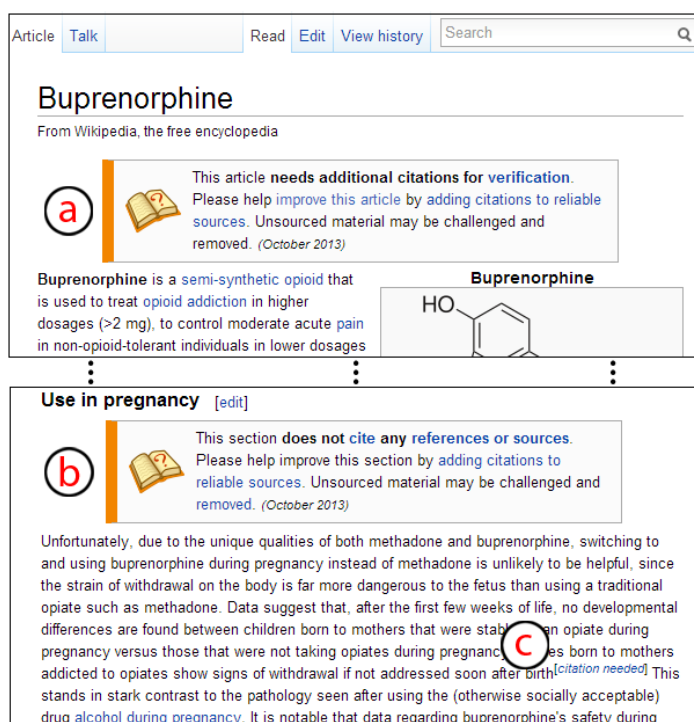


Figure 5.1: Examples for cleanup templates with *a*) page-scope, *b*) section scope, *c*) inline scope
Source of example: <http://en.wikipedia.org/wiki/index.php?oldid=583105148>

usually spelling variations of each other (e.g. POV statement and Pov-statement) or consist of synonymous terms for the same concept (e.g. POV statement and Neutrality disputed).

Similar templates are more difficult to identify. They describe the same quality problem on different levels of granularity or with a different scope. For example, the template POV-title focuses solely on the neutrality of article titles while POV-section or POV indicate neutrality problems in individual sections or whole articles respectively. Even though the target problem of all of these templates is *neutrality*, it will depend on the application whether they should all be aggregated under the same flaw. The similarity of one template to others is often indicated by a reference in the “*See also*” section of the template information page. However, these pages are not structured consistently, which makes automatic extraction of this information impossible. The template category system⁸³ is also not a reliable resource for determining template similarity, since it is merely a functional classification of the templates rather than a well-drafted semantic taxonomy. Therefore, we regard the selection of similar templates to be a manual task.

Finally, unrelated templates are identified by ruling out any similarity or synonymy according to the definition above.

⁸³http://en.wikipedia.org/wiki/Category:Wikipedia_maintenance_templates

For each cluster, one template is defined as the nucleus which is used as the label for the flaw that is represented by the cluster. We choose the template with the highest number of synonyms as the nucleus. In cases of ties, i.e. multiple templates with the same number of synonyms, we choose template with the most comprehensive template information page, since a detailed description indicates a higher importance of the template.

5.2.1.3 Topical Restriction

Many cleanup templates have restrictions concerning the pages they may be applied to. A hard restriction is the page type (or namespace) a template might be used in. For example, some templates can only be used in articles while others can only be applied to discussion pages. This is usually enforced by maintenance scripts running on the Wikimedia servers.

A soft restriction, on the other hand, are the topics of the articles a template can be used in. Many cleanup templates can only be applied to articles from certain subject areas. An example with a particularly obvious restriction is the template `in-universe` (see table 5.3), which should only be applied to articles about fiction. This *topical restriction* is neither explicitly defined nor automatically enforced, but it plays an important role in the quality flaw recognition task, as the remainder of this paper will show.

While flaws merely concerning the structural or linguistic properties of an article are less restricted to individual topics, they are still affected by a certain degree of *topical preference*. Many subject areas in Wikipedia are organized in *WikiProjects*⁸⁴, which have their own ways of reviewing and ensuring quality within their topical scope. Depending on the quality assurance processes established in a WikiProject, different importance is given to individual types of flaws. Thus, the distribution of cleanup templates regarding structural or grammatical flaws is also biased towards certain topics. We will henceforth subsume the concept of topical preference under the term topical restriction.

5.2.2 Definition of the Quality Flaw Detection Task

Based on the above definition of quality flaws, we define the quality flaw detection task similar⁸⁵ to [Anderka et al. \(2012\)](#) as follows:

Given a sample of articles in which each article has been tagged with any cleanup template τ_i from a specific template cluster T_f thus marking all articles in the sample with a quality flaw f , it has to be decided whether or not an unseen article suffers from f .

⁸⁴<http://en.wikipedia.org/wiki/WP:PROJ>

⁸⁵[Anderka et al. \(2012\)](#) consider each flaw to be represented by a single cleanup template rather than by a cluster of similar templates.

| Flaw | κ | F_1 |
|-------------|----------|-------|
| Advert | .60 | .80 |
| Confusing | .60 | .80 |
| Copy-edit | .00 | .50 |
| Essay-like | .60 | .80 |
| Globalize: | .60 | .80 |
| In-universe | .80 | .90 |
| Peacock | .70 | .84 |
| POV | .60 | .80 |
| Technical | .90 | .95 |
| Tone | .40 | .70 |
| Trivia | .20 | .60 |
| Weasel | .50 | .74 |

Table 5.1: Agreement of human annotator with gold standard. The corpus for this small study consist of 20 articles per flaw, half of which are flawed.

We cast this task as a binary classification problem in which a classifier trained on a set of articles that contain the quality flaw f (positive instances) and a set of articles that do not contain f (negative instances) learns to identify unseen articles suffering from f . Therefore, it is both necessary to provide reliable examples and counterexamples for flawed articles in order to achieve a sufficiently high classification performance. However, no articles are marked not to contain a particular quality flaw. Consequently, there is no straight forward way to sample flawless articles for a given flaw f . We therefore propose an approach to extract reliable positive and negative training instances from the article revision history in section 5.3.2. On the data extracted with this approach, we train individual binary classifiers for each quality flaw. It is possible to combine these classifiers in an ensemble method in order to achieve joint classification of multiple flaws (Fujino et al., 2008).

5.2.3 Reliability of Cleanup Templates as Quality Flaw Markers

Arazy and Kopak (2011) discuss that it is important to assess how well humans agree in their quality judgments. This is even more the case when predicting quality flaws that are represented by multiple community-assigned labels. Our approach to quality flaw detection in Wikipedia is based on the assumption that cleanup templates are valid markers of quality flaws. In order to test the reliability of these user assigned templates as quality flaw markers, we carried out an annotation study in which one human annotator was asked to perform the binary flaw detection task manually. For this study, we selected the same set of quality flaws that we define for the NSTYLE corpus described in section 5.3.4. For each flaw, the human rater was provided with the description of the nucleus of the template cluster, which we extracted from the respective template information page. We extracted the plain text of 10 random flawed articles and 10 random untagged articles for each flaw and presented

the texts to the annotator. The annotator had to decide for each flaw individually whether a given text belonged to a flawed article or not. She was not informed about the ratio of flawed to untagged articles.

Table 5.1 lists the chance corrected agreement between the human annotator and the gold standard using Cohen’s κ (Carletta, 1996), which is commonly used to assess agreement between two human raters. The metric is defined as

$$\kappa = \frac{p_0 - p_c}{1 - p_c}$$

where p_0 refers to the observed agreement between the human rater and the gold standard and p_c refers to the chance agreement, which is 0.5 in all cases of this balanced sample corpus. We furthermore report the corresponding F_1 performance for the human predictions against the gold standard defined as follows

$$F_1 = \frac{2 \cdot \text{true positives}}{2 \cdot \text{true positives} + \text{false negative} + \text{false positives}}$$

The templates *copy-edit* and *trivia* yielded the lowest performance in the study. Even though *copy-edit* templates are assigned to whole articles, they refer to grammatical and stylistic problems of relatively small portions of the text. That is, they mark local phenomena rather than the overall state of an article. This increases the risk of overlooking a problematic span of text, especially in longer articles. The *trivia* template, on the other hand, designates sections that contain miscellaneous information that is not well integrated in the article. Upon manual inspection, we found a wide range of possible manifestations of this flaw ranging from an agglomeration of incoherent factoids to well-structured sections that did not exactly match the focus of the article, which is the main reason for the low agreement.

Even though this small scale study is not exhaustive, it gives a clear indication that the scope of cleanup templates has to match the scope of the quality flaw. That is, if a flaw only concerns a small portion of an article, it should be represented by inline- or section scope templates rather than by an article scope template. This issue is not yet addressed in this thesis as it focuses on article classification and thus on article-scope templates alone. However, section 5.6 discusses first attempts towards sentence level classification using cleanup templates. Section 5.7 furthermore considers the general limits of the predictability of quality flaws in Wikipedia.

5.2.4 Coverage of the Article Quality Model by Cleanup Templates

While cleanup templates identify a wide range of quality flaws, not all dimensions defined in the Wikipedia article quality model (see chapter 4.4) are equally well covered. Figure 5.2 gives an overview how well each dimension is represented. The classification is both based

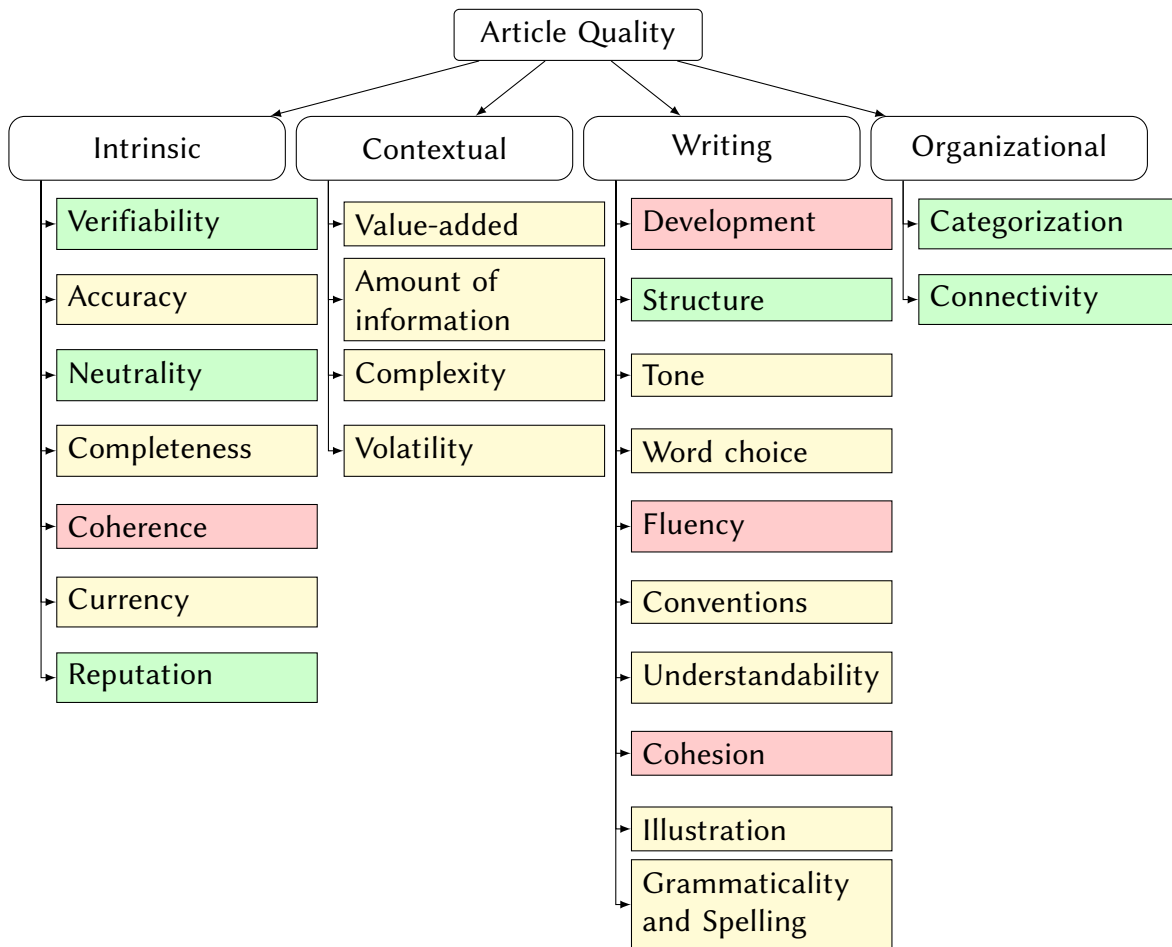


Figure 5.2: Dimensional coverage of the article quality model by cleanup templates. Green indicate full coverage, yellow partial coverage and red denotes inadequate coverage of the dimension.

on the number of different templates that can be assigned to a specific dimension and how much the particular templates are used. We therefore use the cleanup template categorization provided on the Wikipedia cleanup template listing⁸⁶ and manually assign to each category all relevant quality dimensions (see appendix C). Given the number of templates assigned to each quality dimension and the number of occurrences of these templates in Wikipedia, we derive three levels of coverage – *full*, *partial* and *inadequate* – color coded in green, yellow and red. The judgments have been made manually without assigning absolute thresholds of necessary template assignments to each coverage level. Therefore, this categorization is subjective. However, it gives an impression how well the individual dimensions are covered by cleanup templates and for which quality aspects we have to employ other means of assessment.

⁸⁶<http://en.wikipedia.org/wiki/WP:TC>

5.2.5 Quality Flaw Markers in Non-English Wikipedias

Even though this dissertation mainly focuses on the English language and the experiments described in this chapter have been carried out on the English Wikipedia, the methodology itself is language independent. However, in order for the approach to be applicable to a particular language other than English, the respective language edition of Wikipedia needs to make use of a similar system of cleanup templates. It is not easy to exhaustively determine how many of the 287 Wikipedia language versions make active use of cleanup templates as part of their quality assurance process. According to *Wikidata* (see chapter 3.7), 21 language editions have a cleanup template overview page⁸⁷ similar to the English Wikipedia, while 18 Wikipedias have a cleanup template category⁸⁸. When factoring out the overlap between the two, we can assume that at least 31 wikis employ cleanup templates to mark quality flaws.

The English Wikipedia, as the language version with the biggest community, has the most comprehensive system of cleanup templates. However, especially the smaller Wikipedias (e.g. Sinhalese, Orya, Khmer) tend to directly adapt the English system of cleanup templates and therefore exhibit a sophisticated system of flaw markers. Due to the small size of these wikis, the amount of available training data is nevertheless limited.

The German Wikipedia, the second largest language edition in terms of the number of articles, employs a highly reduced set of cleanup templates. They follow the rationale that a large number of tags cannot be handled consistently by a large number of untrained community members and that the reduction to a core selection of templates will therefore be the best solution. The German Wikipedia currently employs 10 templates in four categories. The usage of these templates and the cleanup activities associated with it are centrally monitored on the corresponding WikiProject page⁸⁹.

5.3 Quality Flaw Corpora

In this section, we describe two corpora that we use for our quality flaw detection experiments. The CLEF corpus has been introduced by *Anderka and Stein (2012)* in the *Competition on Quality Flaw Prediction in Wikipedia* as part of the PAN lab at the 2012 Conference and Labs of the Evaluation Forum (CLEF). It is based on the English Wikipedia and represents the ten most common quality flaws.

In our experiments with the CLEF corpus, we found several aspects to negatively influence the reliability and performance of text classifiers trained on this dataset. We therefore

⁸⁷<http://www.wikidata.org/wiki/Q9136874>

⁸⁸<http://www.wikidata.org/wiki/Q8219083>

⁸⁹<http://de.wikipedia.org/wiki/WP:WPWB>

| Flaw | Description | Training | Test |
|-------------------|---|----------|-------|
| Advert | The article appears to be written like an advertisement and should be rewritten from a neutral point of view. | 1 109 | 2 000 |
| Empty section | The article has at least one section that is empty. | 5 757 | 2 000 |
| No footnotes | The article includes a list of references, related reading or external links, but its sources remain unclear because it lacks inline citations. | 3 150 | 2 000 |
| Notability | The article does not meet the general notability guideline. | 6 068 | 2 000 |
| Original research | The article may contain original research and should be improved by verifying the claims made and adding references. | 507 | 1 014 |
| Orphan | The article is an orphan, as no other articles link to it. | 21 356 | 2 000 |
| Primary sources | The article relies on references to primary sources or sources affiliated with the subject and does not contain sufficient citations from reliable and independent sources. | 3 682 | 2 000 |
| Refimprove | The article needs additional citations for verification. | 23 144 | 1 998 |
| Unreferenced | The article does not cite any references or sources. | 37 572 | 2 000 |
| Wikify | The article needs to be wikified, i.e. internal and external links should be added. | 1 771 | 1 998 |
| Untagged | Article without any cleanup templates. | 50 000 | – |

Table 5.2: Flaw definitions and numbers of training and test instances per flaw. The training sets exclusively contain articles tagged with the respective flaw (except for *untagged*). The test sets contain a balanced number of flawed and untagged articles.

discuss these problems in detail and show how they affect machine learning algorithms in real life scenarios.

Finally, we present the NSTYLE corpus, a topically balanced corpus of neutrality and style flaws with reliable negative examples, i.e. documents without the respective flaws, which we designed to solve the problems of the CLEF corpus. It furthermore represents two classes of quality problems that are of particular high importance for the Wikipedia community and are furthermore relevant for textual resources other than Wikipedia.

5.3.1 The CLEF Corpus

The CLEF corpus⁹⁰ reflects the ten most frequently tagged quality flaws in Wikipedia and consists of a training and a test set for each flaw. It is a subsample of the PAN-WQF-12 corpus⁹¹ and has been compiled for the 2012 Competition on Quality Flaw Prediction in Wikipedia (Anderka and Stein, 2012). The training set consists of 104,116 articles extracted from the English Wikipedia snapshot from January 4th, 2012 which are labeled with the respective quality flaws. Furthermore, a set of 50,000 untagged articles is provided. While it is not guaranteed that articles without cleanup tags do not have quality problems, the assumption underlying this corpus is that they provide reasonable examples for articles

⁹⁰Available under <http://www.webis.de/research/events/pan-12/pan12-web/wikipedia-quality.html>

⁹¹<http://www.uni-weimar.de/en/media/chairs/webis/research/corpora/corpus-pan-wqf-12>

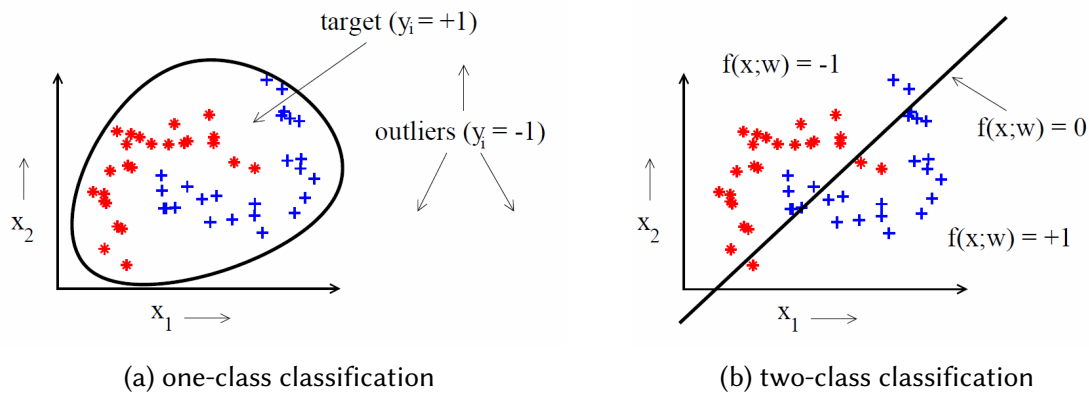


Figure 5.3: Concept of one-class classification according to [Tax \(2001, pp. 4, 14\)](#). One-class classifiers separate all given labeled instances from any outliers, while two-class classifiers separate the two differently labeled classes.

without quality flaws. The test set contains a balanced number of flawed and untagged articles and has a total size of 19,010 documents. Among the untagged articles in the test corpus, 10% are featured articles.

It has to be noted that each flaw in the CLEF corpus is represented by only a single template. Similar templates have not been aggregated to flaw clusters as we suggested in section 5.2.1. We implemented this approach for the NSTYLE corpus that is introduced in section 5.3.4.

Table 5.2 shows the definitions of all flaws in the CLEF corpus as they are displayed on the template information pages and lists the numbers of articles for each flaw in respective training and test set. All articles are provided as plain text in the original MediaWiki markup. In the test set, the cleanup templates representing the quality flaw of the respective class have been removed from the markup, but have been made available as ground truth in a separate file.

5.3.2 Reliability of Training Instances

A central problem of the quality flaw recognition approach based on cleanup template prediction is the fact that no articles are tagged to not contain a particular quality problem. In other words, there are no explicit textual or formal indicators that can be used to retrieve counterexamples for flawed articles. However, the majority of supervised machine learning algorithms for classification problems are two- or multi-class approaches that need both positive and negative examples for learning a decision boundary that is supported from both sides by example instances ([Tax, 2001](#)). So far, two approaches have been proposed by related work to circumvent this problem.

One-Class Classification [Anderka et al. \(2012\)](#) tackle the problem with a one-class classifier that is trained on the positive instances alone thus eradicating the need for negative instances in the training phase (see figure 5.3). [Tax \(2001\)](#) describes three main approaches to one-class classification, i.e. density estimation, boundary methods and reconstruction methods. For all of these approaches, the learner has to be able to measure the distance of any unknown document to the given training examples and has to learn a threshold on the distance for deciding the class assignment. [Anderka et al.](#) use a combination of density estimation and class probability estimation based on the cleanup template frequency in Wikipedia. For evaluating the performance of this classifier in terms of precision, however, it is necessary to provide a set of representative examples that can serve as outliers for the given target class. This closely resembles the initial problem of non-existing negative examples. The authors circumvent the issue by evaluating their classifiers on a set of random untagged instances and a set of featured articles and claim that the actual performance of detecting the quality flaws lies between the two. Therefore, we argue that the original problem is only partially solved by the one-class classification approach.

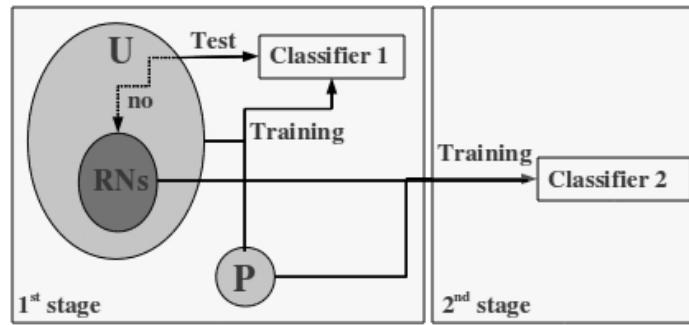
PU Learning [Ferretti et al. \(2012\)](#) follow a two step classification approach designed for learning from positive examples and unlabeled data (PU learning). The idea is to employ different classifiers for preselecting suitable training instances and for performing the actual predictions ([Liu et al., 2002, 2003](#)). In the first phase, the authors use a Naive Bayes classifier trained on positive instances and random untagged articles to pre-classify the data. The assumption behind that is that the negatives identified in the pre-classification step will be better counterexamples than the initially selected random untagged articles. In the second phase, they use these negatives together with the original set of positive instances to train a Support Vector Machine that produces the final predictions. Figure 5.4 shows a schematic overview of the concept. Even though the Naive Bayes classifier is supposed to identify reliable negatives, the authors found no significant improvement over a random selection of negative instances. This, however, effectively renders the PU learning approach redundant.

Since none of the approaches to circumvent the need for negative examples have been effective, we argue that we need to develop a dedicated method for identifying reliable negatives to perform the quality flaw prediction task efficiently and reliably. Our solution to this problem is described in section 5.3.4, where we discuss the construction of our NSTYLE corpus.

5.3.3 Topic Bias

In addition to the reliability of the training instances, another central aspect has to be considered when compiling training corpora for quality flaw classifiers. In section 5.2.1, we discussed that cleanup templates can be topically restricted, i.e. they occur exclusively or

Figure 5.4: Concept of PU learning according to [Ferretti et al. \(2012, p. 4\)](#). Classifier 1 identifies reliable negatives (RNs) among the untagged articles U , which are then used as input for classifier 2. Both classifiers use the same positive instances P .



more likely in articles from particular subject areas. In return, sets of articles that are tagged with the same cleanup template will be biased towards these topics. If this bias is not considered when sampling the negative instances, the positive and the negative set will substantially differ in topic thus resulting in a topically biased dataset. As a consequence, a classifier intended for quality flaw detection is likely to degenerate to a topic classifier and show unrealistically high cross-validated performance in evaluation.

Topic bias is a known problem in text classification. [Mikros and Argiri \(2007\)](#) investigate the topic influence in authorship attribution. They found that even simple stylometric features, such as sentence and token length, readability measures or word length distributions show considerable correlations with the topic. They argue that many features that were largely considered to be topic neutral are in fact topic-dependent variables. Consequently, results obtained on multitopic corpora are prone to be biased by the correlation of authors with specific topics. Therefore, several authors introduce topic-controlled corpora for applications such as author identification ([Koppel and Schler, 2003](#); [Luyckx and Daelemans, 2004](#)) or genre detection ([Finn and Kushmerick, 2006](#)).

[Brooke and Hirst \(2011\)](#) measured the topic bias in the *International Corpus of Learner English* and found that it causes a substantial skew in classifiers for native language detection. In accordance with [Mikros and Argiri](#), the authors found that even non-lexicalized meta features, such as vocabulary size or length statistics, depend on topics and cause cross-validated performance evaluations to be unrealistically high. In a practical setting, these biased classifiers hardly exceed chance performance.

In the context of Wikipedia quality flaw detection, figure 5.5a illustrates the problem that arises from topic agnostic sampling as exhibited by the one-class approach and the PI learning approach⁹² described earlier. Both approaches sample random negative instances A_{nd} for any given set of flawed articles A_f from a set of untagged articles A_u without taking into account the topical restriction for the given flaw f . The articles that conform to the topical restriction of f are indicated by the set A_{topic} that contains articles with a topic

⁹²Even though the PU learning approach selects negative instances with a meta classifier rather than performing random sampling, the result is similar to the random sampling approach as we discussed before. Thus, without loss of generality, we consider the subsample to be random.

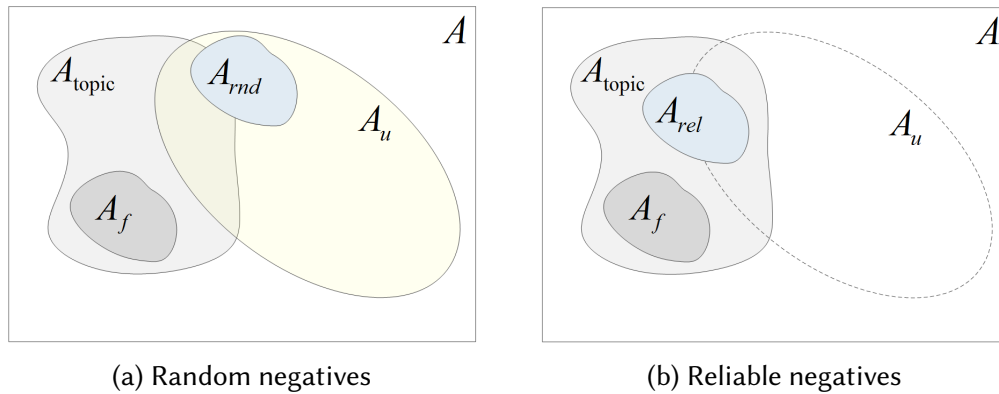


Figure 5.5: Sampling of negative instances for a given set of flawed articles (A_f). Random negatives (A_{rnd}) are sampled from articles without any cleanup templates (A_u). Reliable negatives (A_{rel}) are sampled from the set of articles (A_{topic}) with the same topic distribution as A_f

distribution similar to the flawed articles in A_f . A flaw that is restricted to a very narrow set of topics, such as in-universe, consequently has a small A_{topic} while flaws with a topical preference rather than a topical restriction will have a larger A_{topic} . In any case, the topic distribution of A_{topic} is clearly skewed compared to the near random topic distribution of A_u . Consequently, the topical differences between A_{rnd} and A_f are a predominant feature for a classifier to pick up on.

In order to factor out the article topics as a major characteristic for distinguishing flawed articles from the set of outliers, reliable negative instances A_{rel} have to be sampled from the restricted topic set A_{topic} (see figure 5.5b). This will avoid the systematic bias and result in a more realistic performance evaluation. The fact that it is much easier to determine a decision boundary between A_f and A_{rnd} than between A_f and A_{rel} explains why topic agnostic classifiers show unrealistically high cross-validated results in the evaluation.

In the following section, we describe the NSTYLE corpus in which the topic distributions of negative and positive instances are controlled by dedicated sampling techniques. We furthermore describe our approach how to extract reliable negative and positive training instances from the Wikipedia article revision history.

5.3.4 The NSTYLE Corpus

With the NSTYLE corpus, we pursue two separate goals. First, we extend the scope of the quality flaw detection experiments from the ten most frequent flaws to a newly compiled set of 12 flaws from the categories *neutrality* and *style*. Neutrality is one of the most discussed aspects in the Wikipedia community and directly anchored in the five pillars of Wikipedia (see chapter 3.1). Stylistic aspects, on the other hand, are mainly situated in the writing quality layer of our article quality model and largely underrepresented by existing quality assessment procedures (see section 5.2.4). We therefore chose these two categories to show

that the quality flaw detection approach does not only work for the most frequent flaw types, which was the selection criterion for the flaws in the CLEF corpus. Second, we aim at demonstrating how the influence of the topic bias that was discussed earlier in this section can be factored out by sampling reliable training instances from the revision history.

5.3.4.1 Selection of Flawed Articles

We start with selecting all cleanup templates listed under the categories *neutrality* and *style of writing* in the topology of cleanup templates shown in appendix C. Each of the selected templates serves as the nucleus of a template cluster that potentially represents a quality flaw. To each cluster, we add all templates that are synonymous to the nucleus. The synonyms are listed in the template description under *redirects* or *shortcuts*. Then we iteratively add all synonyms of the newly added template until no more redirects can be found. Furthermore, we manually inspect the lists of similar templates in the *see also* sections of the template descriptions and include all templates that refer to the same concept as the other templates in the cluster. As mentioned earlier, this is a subjective task and largely depends on the desired granularity of the flaw definitions. We finally merge semantically similar template clusters to avoid too fine grained flaw distinctions.

As a result, we obtain a total number of 94 template clusters representing 60 style flaws and 34 neutrality flaws. From each of these clusters, we remove templates with inline or section scope due to the reasons outlined in section 5.2.1.1. We also remove all templates that are restricted to pages other than articles (e.g. discussion or user pages). We use the JWPL (see chapter 3.6.2) to extract all articles marked with the selected templates. We only regard flaws with at least 500 affected articles in the snapshot of the English Wikipedia from January 4, 2012.

Table 5.3 shows an overview of the flaws represented in the NSTYLE corpus. For each flaw, the nucleus of the template cluster is provided along with a description, the number of affected articles, and the size of the template cluster.

5.3.4.2 Extraction of Reliable Instances

As we have argued in section 5.3.2, the extraction of documents that do or do not exhibit a particular quality flaw is an important and non-trivial task. The quality of a machine learning classifier will largely depend on the quality of the data it is trained on and therefore on the success of the data sampling process. In the following, we present our approach to extracting reliable negative training instances that conform with the topical restrictions of the cleanup templates.

Reliable Negatives. Without loss of generality, we assume that an article, from which a cleanup template $\tau \in T_f$ is deleted at a point in time d_τ , no longer suffers from flaw

| | Flaw | Description | Articles | Cluster Size |
|------------|---------------------|---|----------|--------------|
| Neutrality | Advert ^a | The article appears to be written like an advertisement and is thus not neutral | 7,332 | 2 |
| | POV | The neutrality of this article is disputed | 5,086 | 10 |
| | Globalize | The article may not represent a worldwide view of the subject | 1,609 | 1 |
| | Peacock | The article may contain wording that merely promotes the subject without imparting verifiable information | 1,195 | 1 |
| | Weasel | The article contains vague phrasing that often accompanies biased or unverifiable information | 704 | 4 |
| Style | Tone | The tone of the article is not encyclopedic according to the Wikipedia Manual of Style | 4,563 | 6 |
| | In-universe | The article describes a work or element of fiction in a primarily in-universe style ^b | 2,227 | 1 |
| | Copy-edit | The article requires copy editing for grammar, style, cohesion, tone, or spelling | 1,954 | 6 |
| | Trivia | Contains lists of miscellaneous information | 1,282 | 2 |
| | Essay-like | The article is written like a personal reflection or essay | 1,244 | 1 |
| | Confusing | The article may be confusing or unclear to readers | 1,084 | 1 |
| | Technical | The article may be too technical for most readers to understand | 690 | 2 |

^a Also represented in the CLEF corpus.

^b According to the Wikipedia Manual of Style, an in-universe perspective describes the article subject matter from the perspective of characters within a fictional universe as if it were real.

Table 5.3: NSTYLE corpus of neutrality and style flaws. The cluster size refers to the number of templates used to represent the particular flaw (see section 5.2.1)

f at that point in time. Thus, the revision r_{d_t} is a *reliable negative instance* for the flaw f . Additionally, since the article was once tagged with $\tau \in T_f$, it belongs to the same restricted topic set A_{topic} as the positive instances for flaw f .

We use the *Apache Hadoop*⁹³ framework and *WikiHadoop*⁹⁴, an input format for Wikipedia XML dumps, for crawling the whole revision history of the English Wikipedia on a compute cluster to create an index of reliable negative instances for all templates found in the dataset (see figure 5.6). WikiHadoop allows each Hadoop mapper to receive adjacent revision pairs, which makes it possible to compare the changes made from one revision to the next. For every template τ found in the dataset, we extract all pairs of adjacent revisions $(r_{d_{t-1}}, r_{d_t})$, in which the first revision contains τ and the second one does not contain τ , and store them in an aggregated index. From this index, we can retrieve all reliable negative instances for any template.

For extracting the final set of reliable negative instances for a given flaw f , we retrieve from the index all revisions for each template in the template cluster of f . In other words,

⁹³<http://hadoop.apache.org>

⁹⁴<https://github.com/whym/wikihadoop>

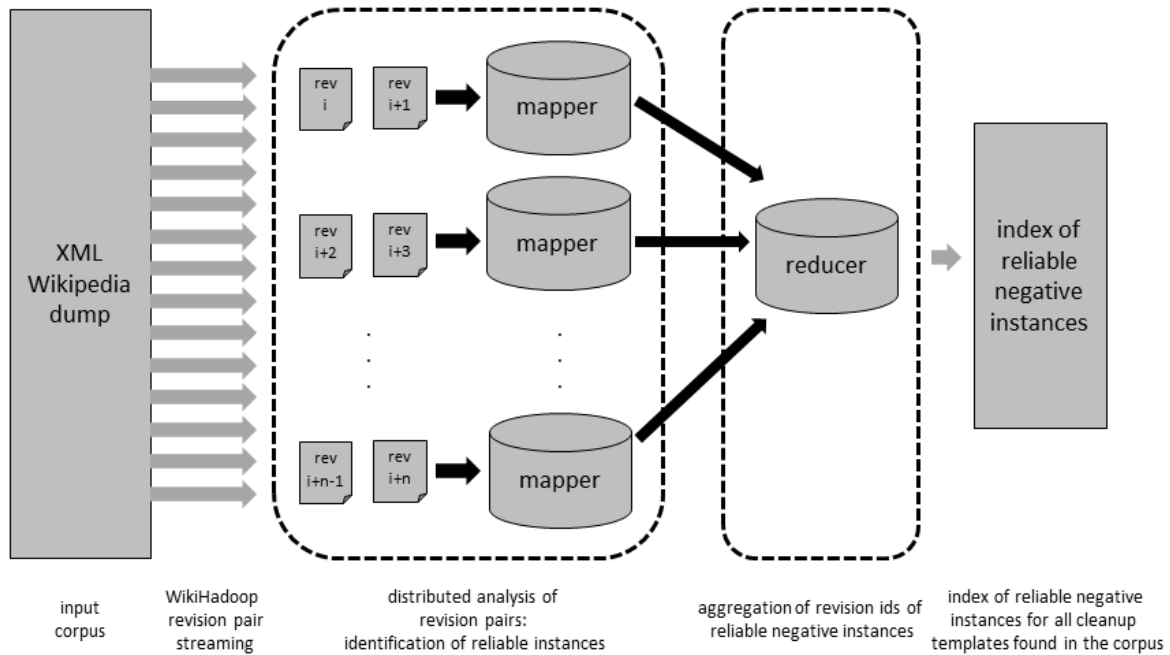


Figure 5.6: Distributed extraction of reliable negative training instances from Wikipedia XML dump on a compute cluster using Hadoop

we retrieve all revisions for any $\tau \in T_f$. Since there are occasions in which a template is replaced by another template from the same cluster rather than being deleted, we ensure that r_{d_τ} does not contain any other template from cluster T_f before we finally add the revision to the set of reliable negatives for flaw f .

The main effort of this approach lies in the one time creation of the index as described above. Creating the actual sets of reliable revisions for individual flaws from this index can be achieved with little work. The performance of the index creation process could further be improved by employing more than one reducer in the MapReduce process. However, in this case, the output of all reducers has to be aggregated again to obtain a single index in the end. We did not attempt this in our experiments. Table 5.5 lists the number of reliable negative instances extracted for each flaw.

Reliable Positives. Even though the issue of finding reliable negative instance is the most immanent, since we do not have any labels indicating articles without a particular problem, it is also important to ensure the quality of the positive instances for which we do have these labels.

The main assumption of the quality flaw detection approach is that Wikipedia articles exhibit a particular quality problem as long as they are marked with the corresponding cleanup template and that the cleanup template is removed as soon as the quality problem is solved. However, as a manual inspection of our corpora has shown, it is possible that quality

problems are solved while the corresponding cleanup templates remain in the article. A scenario for this could be that a user who recently improved an article might first consult the Talk page of the article (see chapter 6) to confirm whether they sufficiently solved the problem before removing the tag. Depending on the discussion activity of the particular article, this could take several days. As a consequence, there is a period of time in which the article carries a cleanup template without exhibiting the corresponding flaw. These instances constitute false positives in the training data.

In order to solve this problem, we have to identify reliable positive instances. Similarly to the approach described above, we backtrack the revision history of every article with a particular cleanup template until we find the revision in which the template first appeared. This revision is regarded as the reliable positive instance.

For our experiments, we only consider pages as positive instances that are marked with a flaw at the time the Wikipedia dump was created. For these pages, we extract the reliable revision as described above. In order to increase the amount of available training data, it is possible to extract additional instances from the revision history to include pages that once suffered from a particular flaw.

Having defined the concepts of reliable negatives and reliable positives, we now describe the three different dataset configurations that we compile from the NSTYLE corpus for our machine learning experiments. The NSTYLE-BASE configuration resembles the sampling methodology of the CLEF corpus and contains the latest revisions of any tagged article as positive instances and random untagged articles as negative instances. The NSTYLE-RELP configuration makes use of the reliable sampling technique for positive instance as described above. As negative instances we again sample random untagged articles. Finally, for the NSTYLE-RELALL configuration, we sample both reliable positives and reliable negatives.

5.3.4.3 Measuring the Topic Bias

As we have argued in section 5.3.3, articles with the same cleanup templates tend to share particular subject areas or topics and can easily be separated from random articles. The more restricted the set of topics of a particular set of positive instances is, the easier it can be separated from random articles with simple lexical features. This topic bias, however, limits the usefulness of the dataset for quality flaw detection experiments, since we are not interested in topic clustering but in identifying quality flaws along with the most descriptive features for these flaws. We therefore first describe a method to quantify the topical similarity between two sets of articles and then measure the similarities between the training sets for the NSTYLE-BASE and the NSTYLE-RELALL configuration in order to show that the topic bias is largely eradicated in the latter approach.

In Wikipedia, the topic of an article is captured by the categories assigned to it. In order to compare two sets of articles with respect to their topical similarity, we represent each

Table 5.4: Cosine similarity scores between the category frequency vectors of the flawed article sets and the respective random or reliable negatives

| Flaw | Cosine Similarity | |
|-------------|-------------------|------------------|
| | (A_f, A_{rel}) | (A_f, A_{rnd}) |
| Advert | .996 | .118 |
| Confusing | .996 | .084 |
| Copy-edit | .993 | .197 |
| Essay-like | .996 | .132 |
| Globalize | .992 | .023 |
| In-universe | .996 | .014 |
| Peacock | .995 | .310 |
| POV | .994 | .252 |
| Technical | .995 | .018 |
| Tone | .996 | .228 |
| Trivia | .980 | .184 |
| Weasel | .976 | .252 |

article set as a category frequency vector. Formally, we calculate for each set the vector $\vec{C} = (w_{c_1}, w_{c_2}, \dots, w_{c_n})$ with w_{c_i} being the weight of category c_i , i.e. the number of times it occurs in the set, and n being the total number of categories in Wikipedia. We can then estimate the topical similarity of two article sets by calculating the cosine similarity of their category frequency vectors $\vec{C}_1 := A$ and $\vec{C}_2 := B$ as

$$\text{sim}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

Table 5.4 gives an overview of the similarity scores between each positive training set and the corresponding reliable negative set as well as between each positive set and a random set of untagged articles.

We can see that the topics of articles in the positive training sets are highly similar to the topics of the corresponding reliable negative articles while they show little similarity to the articles in the random set. This implies that the systematic bias introduced by the topical restriction has largely been eradicated by our approach.

Individual flaws have differently strong topical restrictions. The strength of this restriction depends on the size of A_{topic} , i.e. the set of articles with the same topic distribution as the flawed articles. In other words, a flaw such as *in-universe* is restricted to a very narrow selection of articles, while a flaw such as *copy edit* can be applied to most articles and rather shows a topical preference due to reasons outlined in section 5.2.1. It is therefore to be expected that flaws with a small A_{topic} are more prone to the topic bias.

5.3.4.4 Corpus Analysis

We close this section with an overview of the properties of the NSTYLE corpus. Table 5.5 lists the number of reliable positive and negative instances for each flaw type. It furthermore shows the total number of tagged revisions in the snapshot of the English Wikipedia from January 4, 2012 and the date of the first appearance of each flaw. The latter has been computed by identifying the oldest article revision in the Wikipedia dump that contained any cleanup template from the template cluster of the respective flaw.

While the flaws Technical and Weasel hardly exceed the minimum of 500 affected articles that we defined in the sampling process, the majority of flaws exhibit between 1,000 and 2,000 affected articles. The most frequently observed flaws, Advert, POV and Tone, occur nearly 17,000 times – more often than all the other nine flaws combined.

The number of reliable negatives differs more substantially across all flaw types depending on how actively a particular set of cleanup templates has been used, how fast the flaw can be corrected and how long the type of flaw already exists in the English Wikipedia. While the number of positive instances indicates the current⁹⁵ articles that suffer from this flaw, the number of negatives indicates the total number of times the particular flaw was corrected. We can therefore see the ratio of positives to negatives as a rough proxy for how easy, and potentially how fast, a particular flaw can be corrected. For instance, 5,086 articles were marked with the flaw POV at the time the Wikipedia dump was created while cleanup templates belonging to the POV template cluster have been removed from 105,066 articles in the past. In contrast, the flaw Advert occurred 7,332 times at time of dump creation, but was only corrected 39,133 times before. Even though the Advert flaw appeared for the first time roughly one year after the POV flaw, these numbers allow the assumption that POV flaws are corrected faster than Advert flaws. [Anderka \(2013\)](#) performed an analysis of average correction times for cleanup templates without considering their aggregation to template clusters. He found that the average time needed to fix an article scope template is 176 days.

We provide descriptive statistics for the flawed articles in the NSTYLE corpus in figure 5.7 including an overview of the article age, number of unique contributors, number of revisions and the article length in tokens. This should give an impression of the properties of the articles in the corpus rather than characterize the individual quality flaws.

5.4 A System for Quality Flaw Detection

In this section, we first describe the architecture and setup of FlawFinder, our quality flaw prediction system, and then provide an overview of the features used to capture quality flaws. In the following section, we furthermore proceed with a description of our experi-

⁹⁵At the point in time when the Wikipedia dump was created.

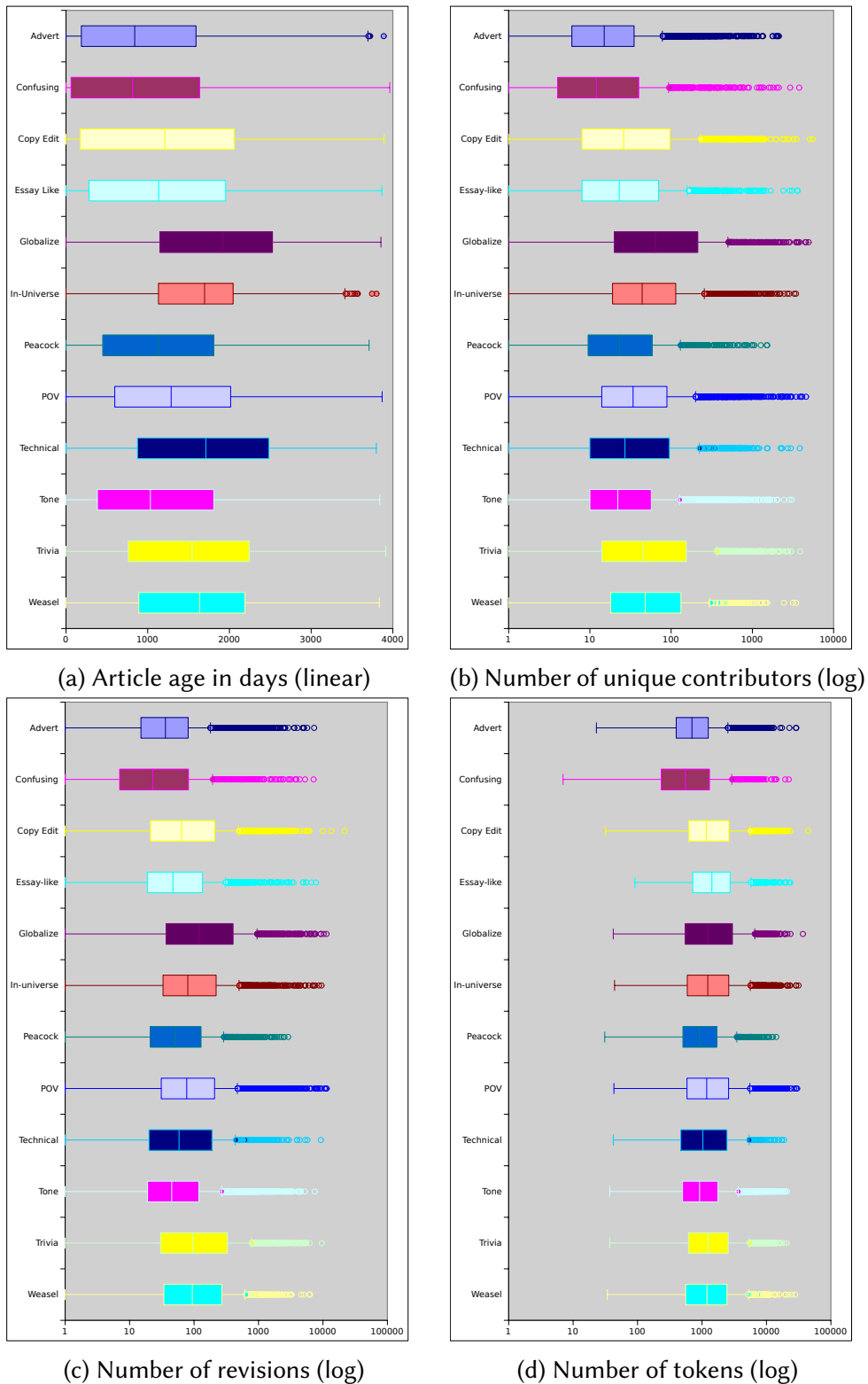


Figure 5.7: Descriptive statistics for the *flawed* articles in the NSTYLE corpus. The article age is displayed on a linear scale while the other properties are plotted on a logarithmic scale.

| Flaw | Positives | Negatives | Total Revisions | First Appearance |
|-------------|-----------|-----------|-----------------|------------------|
| Advert | 7,332 | 39,133 | 627,844 | 2005-06-06 |
| Confusing | 1,084 | 6,225 | 208,296 | 2005-03-20 |
| Copy-edit | 1,954 | 2,878 | 168,423 | 2004-12-30 |
| Essay-like | 1,244 | 3,898 | 164,243 | 2007-04-23 |
| Globalize | 1,609 | 8,196 | 439,264 | 2005-09-08 |
| In-universe | 2,227 | 5,270 | 332,159 | 2006-06-20 |
| Peacock | 1,195 | 7,022 | 169,199 | 2006-02-19 |
| POV | 5,086 | 105,066 | 2,442,626 | 2004-05-31 |
| Technical | 690 | 2,056 | 77,518 | 2005-02-25 |
| Tone | 4,563 | 20,166 | 948,227 | 2005-01-01 |
| Trivia | 1,282 | 70,304 | 2,601,217 | 2005-04-13 |
| Weasel | 704 | 12,710 | 397,238 | 2005-10-07 |

Table 5.5: This table lists the number of positive and negative instances per quality flaw in the NSTYLE corpus. The column *total revisions* furthermore lists the number of revisions in the Wikipedia snapshot from January 4, 2012 that are tagged with any cleanup template from the template cluster of the respective flaw. The *first appearance* refers to the timestamp of the oldest article revision in the dump that contains a template from the respective cluster.

ments both on the CLEF corpus⁹⁶ and the NSTYLE corpus, followed by a detailed evaluation and error analysis.

5.4.1 System Architecture

FlawFinder has been implemented as a modular and highly flexible text classification system based on the Unstructured Information Management Architecture (UIMA) (Ferrucci and Lally, 2004). Even though FlawFinder has been developed to predict quality flaws in unseen texts, its basic design can be used for generic text classification tasks. In fact, the system has been further developed into a generic system for supervised learning on textual data and made publicly available as the DKPro Text Classification Framework (DKPro TC) on Google Code (Daxenberger et al., 2014). This general purpose framework is further described in appendix A.2.

The component software architecture of UIMA enables applications that implement this framework to be decomposed into reusable components that can be arranged into processing pipelines. Within these processing pipelines, the documents are passed on as a common analysis structure (CAS) that can be consumed by every downstream component. These components do not alter the document directly, which remains immutable throughout the

⁹⁶Since the CLEF corpus does not define template clusters, we regard the provided cleanup template to represent a single-element cluster. Thus, the same task definition can be used for this dataset.

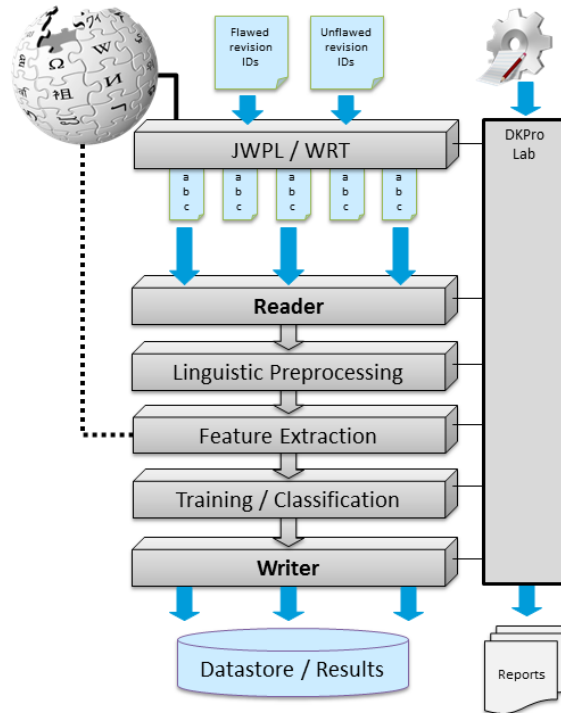


Figure 5.8: High-level system architecture of the FlawFinder

process, but any analysis output or generated information is rather stored in the CAS as a standoff annotation.

For additional flexibility and modularity, we employ the DKPro Lab (Eckart de Castilho and Gurevych, 2011) as a runtime environment for FlawFinder. The DKPro Lab is a light-weight framework that allows to combine independent NLP pipelines into one integrated and highly configurable task-based system. Each task is a self-sufficient processing unit containing a single UIMA pipeline and is responsible for its own data management. Configuration parameters can be injected into each task, whereas the results of the task with each configuration are stored and re-used whenever possible. Furthermore, it is possible to attach reports to each task in order to monitor, summarize or post-process the intermediate task output or final experiment results. The main advantage of the DKPro Lab is its parameter sweeping functionality. That is, by providing value ranges for each parameter an experiment depends on, the DKPro lab handles the parameter combinations that have to run and reuses the intermediate output that has already been calculated in an earlier configuration.

Overall, FlawFinder consists of five main components, a corpus reader, a linguistic preprocessing engine, a feature extraction unit, a module for training and evaluating classification models, and a report writer.

Corpus Reader. FlawFinder has been designed as a binary, single label text classification system. That is, a single run of the system always focuses on an individual flaw f and learns how to determine whether an article suffers from this flaw. Thus, the corpus reader needs to provide the corresponding training instances that are marked with f , i.e. positive instances, and instances that do not exhibit f , i.e. negative instances.

For the sake of flexibility, our corpora do not consist of full text but merely of IDs linking to the corresponding articles or article revisions in a preprocessed Wikipedia database. The database is created and accessed with JWPL (section 3.6.2), while access to the revision history is provided by the Wikipedia Revision Toolkit (see appendix A.1). This way, we are able to obtain any available metadata later on without having to decide in advance about what information to include in the corpora and how to structure the data.

The corpus readers are fed with lists of IDs that have been selected with the sampling techniques described in section 5.3. Depending on the configuration of the experiment, these IDs either refer to articles or particular article revisions (reliable training instances). Each experiment consists of two sets of IDs containing references to flawed and flawless articles respectively. The reader loads the articles from the database into the processing pipeline, marks them with the flaw label and any necessary credentials for accessing the article in the database. This way, any downstream component that merely requires the article text can directly work on the document that passes through the pipeline while any component demanding additional information from the database, such as the number of pages linking to the particular article, can access this information later on (see figure 5.8).

Linguistic Preprocessing. The main goal of the linguistic preprocessing module is to prepare the documents for later processing by the feature extractors. It uses NLP components from DKPro Core (Gurevych et al., 2007) for sentence splitting, tokenization, stop word annotation and named entity recognition and can be extended with additional components depending on the requirements of the feature extractors. We furthermore use the SWEBLE parser (Dohrn and Riehle, 2011a) for parsing the wiki markup and creating a Wikitext Object Model (WOM) representation of the article (Dohrn and Riehle, 2011b). This WOM is an abstract syntax tree that serves the same purpose as a document object model (DOM) for an HTML document. It can be used to query the content of an article in a structured manner.

Feature Extraction. The feature extraction module has been implemented using ClearTK (Ogren et al., 2008), a UIMA-based framework for developing statistical NLP components. It offers interfaces for creating feature extractors that can be used independently from the utilized machine learning algorithm. Even though the extractors are UIMA components and thus can consume annotations created by the preprocessing module, they do not store their output as annotation in the CAS. They rather pass the extracted features to a central feature store which is maintained by the ClearTK framework. When the extraction process

is finished for the whole document collection, the contents of the feature store are converted into the particular format that is required by the machine learning components used in the setup. This decoupling of feature extraction mechanics and machine learning specific formatting makes it possible to create a highly configurable and modular feature extraction pipeline without imparting restrictions on the downstream components for training and classification.

Machine Learning. For training the classifiers and evaluating their performance, the machine learning module employs two different machine learning toolkits. For the experiments on the CLEF corpus we use Mallet, the *Machine Learning for Language Toolkit* (McCullum, 2002), since it offers a small collection of widely used text classification algorithms and is directly supported by the ClearTK framework. For the experiments in the NSTYLE corpus, we employ Weka (Hall et al., 2009), a Java-based data mining toolkit that provides a larger selection of machine learning algorithms.

Reporting. The reporting module gathers information from the other tasks and generates both individual, detailed reports for each experiment configuration and an overall report with a summary of the results of all configuration runs in a particular setup. This makes it possible to easily compare the performance of different classifiers and classification parameters.

5.4.2 Features

In order to capture the ten quality flaws represented in the CLEF corpus, we initially defined a set of 29 feature types that – according to a manual inspection of flawed articles and according to the flaw definitions – most likely indicate the presence or absence of any of the selected quality flaws. Consequently, these features are very specific to the flaws they are supposed to predict.

For the experiments on the NSTYLE corpus, we aimed at finding universal features that indicate style and neutrality issues rather than tailoring particular features to detect single flaw types. Furthermore, in order to gain insights how individual feature categories perform on detecting style and neutrality flaws, we grouped the features into four feature sets. NSTYLE-NONGRAM excludes all lexical features while NSTYLE-NGRAM is restricted to lexical features. NSTYLE-NOWIKI excludes all wiki-specific features such as markup, link structures or categories. We compiled this set in order to identify textual characteristics that can be transferred to texts other than Wikipedia articles. Finally NSTYLE-ALL includes all features relevant for the NSTYLE corpus without any additional restrictions.

In the remainder of this section, we describe the features used in the CLEF and NSTYLE experiments. An overview of all features and how they are combined in the classification experiments is shown in table 5.6.

Structural Features are supposed to capture basic structural properties and surface characteristics of the Wikipedia articles. As described in the system architecture, we use the SWEBLE parser to create a Wikitext Object Model (WOM) of each page. From this model, we extract all article sections along with their headers. We use the number of sections, the mean length of the section texts and the number of empty sections as features. Furthermore, we extract a plain text representation without wiki markup from the WOM and calculate the ratio of markup to plain text as a fourth surface feature.

Reference Features capture aspects regarding the use of citations in the article. There are basically two types of references, *footnote style* references and *bibliography style* references. Footnote style references are marked with `<ref> ...<\ref>` tags directly in the text and are automatically listed at the bottom of the page⁹⁷. Bibliography style references are manually listed at the end of the article, usually in the *References* section. They can either be created as manually formatted list items or can be marked with `cite` or `citation` tags for automatic reference formatting. First, we check whether manually created bibliography items exist in the *References* section and how many elements it contains. Then we count the number of all inline references in the article and determine their average number per sentence. Finally, we determine the ratio of the number of all references to the length of the article. Analogously to lists of references, it is possible to define lists of explanatory notes using the `{{notelist}}` template. It is usually placed in the *Notes* section and gathers all occurrences of explanatory notes which are defined within the text with `efn` templates. We extract this information in the same way as the references.

Network Features reflect the connections of an article within the whole network of Wikipedia articles and to external resources. Since the number of inbound links (i.e. the number of times other articles link to a given article) cannot be determined by parsing the articles in the provided corpora alone, we use the respective information from our JWPL Wikipedia database. When creating a new Wikipedia database from a Wikipedia data dump, JWPL automatically parses the articles using the JWPL Wikitext parser and stores the link information in the database. For each article, we determine the number of wiki-internal inbound links, wiki-internal outbound links and links to resources outside of Wikipedia.

⁹⁷Depending on the setup of the page, the references might appear in different sections such as *References*, *Notes* or *Citations*.

Table 5.6: Feature sets used in the experiments on the CLEF and NSTYLE corpora.

indicates numbers of instances

| Category | Feature type | CLEF-ALL | NSTYLE-ALL | NSTYLE-NONGRAM | NSTYLE-NGRAM | NSTYLE-NOWIKI | NSTYLE-ALL |
|--------------|--------------------------------------|----------|------------|----------------|--------------|---------------|------------|
| Lexical | Article ngrams | • | • | • | • | | |
| | Info to noise ratio | | • | | • | | |
| Network | # External links | • | • | | | • | |
| | # Inlinks | • | | | | | |
| | # Inlinks<3 | • | | | | | |
| | No inlinks | • | | | | | |
| | # Outlinks | • | • | | | | • |
| | # Outlinks per sentence | | • | | | | • |
| | # Language links | | • | | | | • |
| References | Has reference list | • | • | | | • | |
| | # References | • | • | | | • | |
| | # References per sentence | • | • | | | • | |
| | References to text ratio | • | | | | | |
| Revision | Has references | • | | | | | |
| | # Revisions | • | • | | | | • |
| | # Unique contributors | • | • | | | | • |
| | # Registered contributors | • | | | | | |
| Structure | Article age | • | | | | | |
| | # Empty sections | • | • | | | | • |
| | Mean section size | • | • | | | | • |
| | # Sections | • | • | | | | • |
| | # Lists | | • | | | | • |
| | Question rate | | • | | • | | • |
| | Markup to text ratio | • | | | | | |
| Readability | ARI | | • | | | • | • |
| | Coleman-Liau | | • | | | • | • |
| | Flesch | | • | | | • | • |
| | Flesch-Kincaid | | • | | | • | • |
| | Gunning Fog | | • | | | • | • |
| | Lix | | • | | | • | • |
| | SMOG-Grading | | • | | | • | • |
| Named Entity | # Person entities | • | • | | | • | • |
| | # Organization entities | • | • | | | • | • |
| | # Location entities | • | • | | | • | • |
| | # Person entities per sentence | • | | | | | |
| | # Organization entities per sentence | • | | | | | |
| | # Location entities per sentence | • | | | | | |
| Misc | # Characters | • | • | | | • | • |
| | # Sentences | • | • | | | • | • |
| | # Tokens | • | • | | | • | • |
| | Average sentence length | | • | | • | | • |
| | Article lead length | | • | | | | • |
| | Lead to article ratio | | • | | | | • |
| | # Discussions | • | • | | | | • |

Named Entity Features capture the number of named entities in the article. We use the Stanford Named Entity Recognizer (Finkel et al., 2005) using the 3-class model with distributional similarity features⁹⁸ for tagging all entities of the types Person, Organization and Location. We use both the overall named entity counts and the average number of named entities per sentence as features.

Revision-based Features are based on metadata derived from the article revision history. We use the Wikipedia Revision Toolkit (WRT) (Ferschke et al., 2011) to determine the number of revisions for each article. Furthermore, we count the number of unique users that edited the page in the past. Since this number also includes anonymous users, which might be counted several times due to changing IP addresses, we additionally determine the number of unique registered users. Finally, we capture the age of the article in days. The WRT is described in more detail in appendix A.1.

Lexical Features are extracted from the plain article text that we obtain from the Wikitext Object Model created by the SWEBLE parser. Any wiki markup is removed except for internal and external links. All links are replaced with a generic EXPLICITLINK label. Furthermore, we perform stopword filtering using the stopword list from the snowball stemmer⁹⁹, which we augmented with punctuation marks. We extract all token-unigrams, bigrams and trigrams from each article and disregard any ngrams with a frequency lower than 5 across the corpus. This cutoff value was determined empirically during the parameter optimization run. We found that a value of 5 was optimal for all flaws.

Readability Features measure the clarity of writing and the level of reading competency needed to understand a text. Most of the prominent metrics rely on surface features that consider average word and sentence length along with the number of syllables per sentence. We use the metrics implemented in the readability package of DKPro Core (Gurevych et al., 2007) including the Flesch-Kincaid grade level metric (Kincaid et al., 1975), the Automatic Readability Index (ARI) (Smith and Senter, 1967), the LIX index (Björnsson, 1968), the Coleman-Liau index (Coleman and Liau, 1975), the Flesch reading ease test (Flesch, 1948), the SMOG grade metric (McLaughlin, 1969) and the Gunning-Fog index (Gunning, 1969).

Other Features include character counts, token counts and sentence counts per article. Furthermore, we measure the discussion activity by means of counting the number of individual discussion topics on the Talk page associated with the article. According to Ferschke et al. (2012a), we regard each titled section on the Talk page as a discussion topic. We refrain from using lexical features from Talk pages, since the Talk page could explicitly discuss the

⁹⁸<http://nlp.stanford.edu/software/CRF-NER.shtml>

⁹⁹<http://snowball.tartarus.org/algorithms/english/stop.txt>

cleanup tags that are supposed to be predicted. This information leak would thus lead to biased results.

5.5 Experiments

In this section, we first describe the experiment setup on both corpora and how we optimize the experiment configuration followed by an evaluation of the classifier performance and an error analysis.

5.5.1 Experiment Setup and Optimization

The experiments on the CLEF and the NSTYLE corpora have both been carried out with the FlawFinder system using different machine learning toolkits for training and evaluating the classifiers. In particular, we switched from the Mallet machine learning toolkit (McCallum, 2002) to Weka (Hall et al., 2009) due to its wider range of machine learning algorithms and easier to use data format. This section outlines the setup of each experiment and how we determined the best configuration for the final evaluation.

5.5.1.1 CLEF Experiment Setup

For the experiments on the CLEF corpus, we use two machine learning algorithms from the Mallet machine learning toolkit (McCallum, 2002), a *Naive Bayes* classifier and *C4.5 decision trees*. For efficiently training the Naive Bayes classifier, we perform unsupervised discretization of numeric features using equal interval binning as suggested by Witten et al. (2011), since the algorithm does not cope well with real valued features and the Mallet toolkit is not able to perform feature discretization automatically. The decision trees were trained using adaptive boosting with 100 rounds and were limited to the depth of five due to memory restrictions.

We experimentally derived the best configuration for each flaw in a *parameter optimization run*, which consists of several training iterations on the same training subset using different parameters. To this end, we evaluate the performance of both algorithms for each flaw on 10-fold cross validation using 500 positive and 500 negative instances from the training set. We parameterize each run with the number of selected features (between 250 and 1,500), the use of a stop-word filter and the frequency cut-off for discarding rare ngrams in order to obtain the best setting.

We use the Information Gain feature selection approach (Mitchell, 1997) to rank and prune the feature space. Table 5.7 shows the result of the feature selection process for each flaw and lists the selected features along with their feature utility scores. The scores depict the discriminativeness of each feature for a given flaw and are the basis for the

feature ranking we derived during training. This information sheds light on which types of features work best to represent the individual flaws. A detailed evaluation of the classifier performance along with an error analysis will be provided in section 5.5.2.

It is not surprising that the best indicators for structural flaws are the corresponding structural properties, such as *has empty section* for *Empty Section*. For other flaws, the feature ranking is more interesting. For *Original Research*, for instance, the best ranked feature is the discussion activity. This suggests that the discussion content might also be informative for identifying this flaw and that the Talk pages should be further exploited for feature extraction. For the flaw *Advert*, the most discriminative non-lexical features are links pointing to external resources. Taking into account the content to which these external links point could further improve the classification performance. It has to be noted that the utility scores cannot be directly compared across flaws. They are only significant as indicators for the ranking within a given flaw. Lexical features are most effective for the flaws *Advert*, *Notability* and *Original Research*, while the other flaws only show little performance gain when adding ngrams to the feature sets. This is to be expected, since structural flaws such as *Empty Section* or *Wikify* are not expressed by the vocabulary but by the article structure and the markup.

5.5.1.2 NSTYLE Experiment Setup

While the experiments on the NSTYLE corpus are based on the same system as the CLEF experiments, we made minor adjustments to the experiment setup. Mainly, we replaced the Mallet machine learning toolkit with Weka in order to gain access to a larger collection of classification algorithms.

We furthermore adapted our feature selection approach to the two step strategy that was able to improve the time efficiency of the parameter estimation process. We first filter the ngrams according to their document frequency in the training corpus. We discard all ngrams that occur in less than $x\%$ and more than $y\%$ of all documents. Several values for x and y have been evaluated in parameter tuning experiments. The best results have been achieved with $x=2$ and $y=90$. In a second step, similar to the CLEF setup, we apply the Information Gain feature selection approach to the remaining set to determine the most useful features.

As we have discussed in the corpus description, we employ three different dataset configurations derived from the NSTYLE corpus. The NSTYLE-BASE configuration uses the newest version of each flawed article as positive instances and a random set of untagged articles as negative instances. The NSTYLE-RELP configuration uses reliable positives, as described in section 5.3.2, in combination with random outliers. Finally, the NSTYLE-RELALL configuration employs reliable positives in combination with the respective reliable negatives.

| | Advert | Empty Section | Notability | Original Research | Refimprove | Unreferenced | Orphan | Wikify | No Footnotes | Primary Sources | |
|-------------------------------------|-------------|---------------|------------|-------------------|------------|--------------|--------|-------------|--------------|-----------------|--------------|
| Selected features | 1,500 | 500 | 250 | 250 | 250 | 1,000 | 1500 | 1,500 | 250 | 1,500 | |
| Classifier | NB | C45 | NB | NB | NB | C45 | C45 | NB | C45 | NB | |
| Selected unigrams | 772 | 126 | 84 | 238 | 0 | 0 | 855 | 769 | 91 | 860 | Lexical |
| Selected bigrams | 602 | 210 | 97 | 2 | 155 | 639 | 399 | 646 | 90 | 404 | |
| Selected trigrams | 116 | 161 | 68 | 0 | 82 | 350 | 229 | 364 | 56 | 223 | |
| #Revisions | .008 | | | .020 | | | | .008 | | .006 | Revision |
| #Contributors | .015 | | | .020 | .013 | | | .016 | | | |
| #Registered contributors | .015 | | | .023 | | | | .018 | | | |
| Article age | .007 | | | | | | | .012 | | | |
| Has empty section | | .534 | | | .004 | .007 | .002 | | | | Surface |
| Markup to text ratio | | | | .017 | .001 | .003 | | .003 | | .003 | |
| Mean section length | | .034 | | | | .022 | .002 | .005 | .025 | | |
| #Sections | .010 | .033 | | | .018 | .023 | .002 | | .025 | .014 | |
| #References | | | | | .029 | .250 | .006 | .003 | .071 | .004 | Reference |
| #References per sentence | | | | | .002 | .250 | .006 | | .071 | | |
| References to text ratio | | | | .017 | | .250 | .006 | | .071 | | |
| Has references | | | | | | | | | | | |
| Has reference list | | | | | | .013 | | .003 | | | |
| #External links | .067 | | | | .007 | .097 | .001 | | .026 | .050 | Network |
| #Inlinks | | | | | | | .145 | .004 | .005 | .003 | |
| #Outlinks | .013 | | | .002 | | | | .011 | | .007 | |
| Inlinks<3 | | | .045 | .025 | | | .472 | .069 | | .002 | |
| No inlinks | | | | | | | .145 | | .005 | | |
| #Organization entities | | | | | | | | | .015 | | Named Entity |
| #Person entities | | | | | | | | | .006 | | |
| #Location entities | | | | | | | | | | | |
| #Organization entities per sentence | | | | | | | | | .015 | | |
| #Person entities per sentence | | | | | | | .002 | .003 | .006 | | |
| #Location entities per sentence | | | | | | | .002 | .005 | | .004 | |
| #Discussions on Talk page | .024 | | | .144 | .048 | | .018 | .010 | .016 | .005 | Other |
| #Characters | .021 | | | .031 | .005 | .003 | .003 | .008 | | .012 | |
| #Sentences | | | | .025 | .005 | .003 | .003 | .004 | | .005 | |
| #Tokens | .021 | | | .031 | | | .003 | .009 | | .009 | |

Table 5.7: Overview of the feature utility scores (information gain) of non-lexical features per quality flaw on the CLEF corpus. The highest ranked feature for each flaw is written in bold. Missing values indicate that the feature has not been selected by the feature selector. The values for lexical features are numbers of selected ngrams per feature type.

| Algorithm | Average F_1 |
|--------------------------------|---------------|
| SVM RBF Kernel | 0.82 |
| AdaBoost (decision stumps) | 0.80 |
| SVM Poly Kernel | 0.79 |
| RBF Network | 0.78 |
| SVM Linear Kernel | 0.77 |
| SVM PUK Kernel | 0.76 |
| J48 | 0.75 |
| Naive Bayes | 0.72 |
| MultiBoostAB (decision stumps) | 0.71 |
| LibSVM One Class | 0.67 |
| Logistic Regression | 0.60 |

Table 5.8: Average F_1 -scores over all flaws on NSTYLE-RELP using NSTYLE-ALL features

While we restricted the experiments on the CLEF corpus to two classifiers from the Mallet toolkit, we explored a wider range of learning algorithms from the Weka toolkit that are known to work well in similar tasks in order to assess their suitability for quality flaw detection on the NSTYLE corpus. This exploratory evaluation was carried out on the NSTYLE-RELP configuration using all available features. A list of all learning algorithms along with the average F_1 -score achieved on NSTYLE-RELP is shown in table 5.8. The performance has been evaluated with 10-fold cross validation on 2,000 documents split equally into positive and negative instances. One class classifiers are trained on the positive instances alone. We determined the best parameters for each algorithms in a parameter optimization run and only list the results of the best configuration.

Overall, Support Vector Machines with RBF kernels yielded the best average results and outperformed the other algorithms on every flaw. We used a sequential minimal optimization (SMO) algorithm (Platt, 1998) to train the SVMs and used different γ -values for the RBF kernel function. In contrast to Ferretti et al. (2012), we did not see significant improvements when optimizing γ for each individual flaw, so we determined one best setting for each dataset. Since SVMs with RBF kernels are a special case of RBF networks that fit a single basis function to the data, we also used general RBF networks that can employ multiple basis functions, but we did not achieve better results with that approach.

One-class classification, as proposed by Anderka et al. (2012), did not perform well within our setup. Even though we used an out-of-the-box one class classifier, we achieve similar results as Anderka et al. in their pessimistic setting, which best resembles our configuration. However, the performance still lacks behind the other approaches in our experiments. The best performing algorithm on the CLEF corpus, AdaBoost with decision stumps as a weak learner, showed the second best results in the exploratory evaluation on NSTYLE.

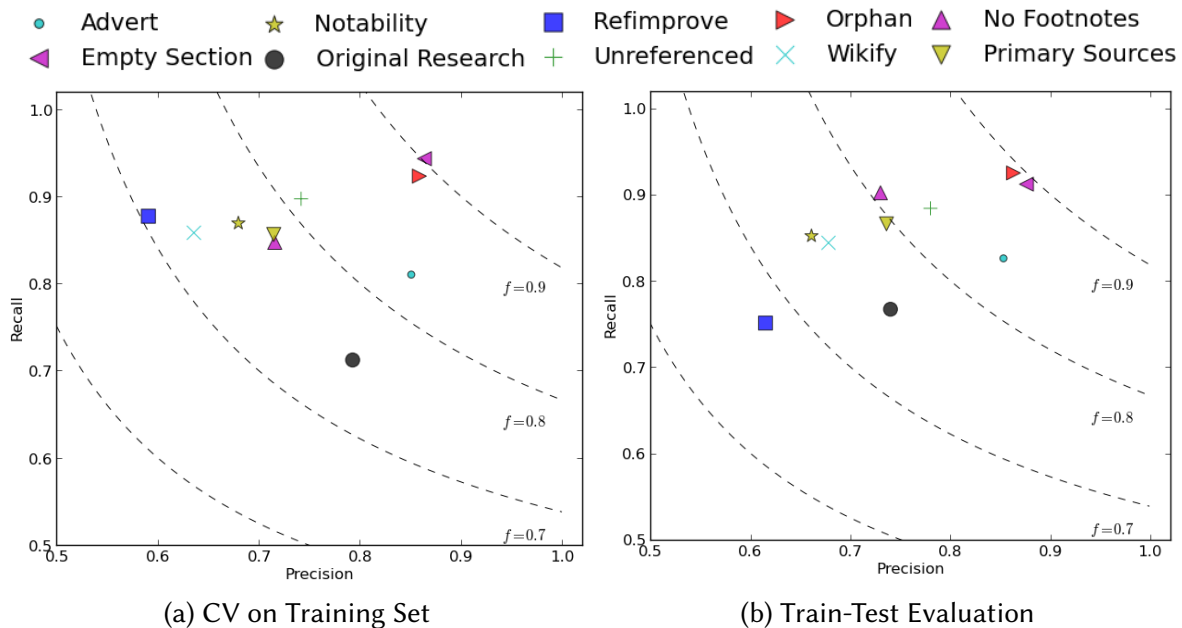


Figure 5.9: Classifier performance on CLEF in terms of precision, recall and F_1 -score.

5.5.2 Evaluation and Error Analysis

In this section, we separately evaluate the performance of the final classifiers trained on CLEF and NSTYLE with the best configuration we derived in the parameter optimization phase as discussed above and analyze the systematic errors made by the classifier.

5.5.2.1 CLEF

Figure 5.9 shows an overview of the classification performance on the training and test set. Figure 5.9a shows the results on the training data with the best configuration obtained in the parameter optimization run and derived in a 10-fold cross validation. Figure 5.9b shows the performance of the final classifiers trained on the whole training data and evaluated on the test data which was compiled by the organizers of the quality flaw prediction competition.

The good performance on the *Advert* flaw comes surprising, since it initially seemed to be a hard task due to the subjectiveness and subtlety of this flaw. The lexical features are good indicators for the presence of this flaw. The most highly ranked ngrams mainly consist of references to business and industry, which can be seen in this list of the top ten ngrams for this flaw:

companies, 's, based, business, company, offers, based in, management, services, products

The selected lexical features are thus highly relevant for the advert context and very predictive of the flaw. On the other hand, the relatively weak performance on *Wikify* was

not expected. The prediction of this flaw particularly suffered from the selection of negative instances in the training set to which we proposed a solution earlier and which we demonstrate on the NSTYLE corpus. We furthermore found that the categories *Original Research*, *Refimprove* and *Primary Sources* have fuzzy boundaries and that Wikipedians do not use these flaw markers consistently. They are often confused with each other, which results in biased training data. Cleanup tags related to references and citations should be consolidated into fewer labels with distinct boundaries.

Compared to the competing systems, FlawFinder achieved the second best results on the CLEF corpus in terms of overall F_1 -score. We argue that the precision of a quality flaw classifier is more important than its recall because it is supposed to facilitate the human review of articles by listing the most likely candidates suffering from particular flaws. With respect to precision, FlawFinder achieved the best results on seven out of ten flaws. This is also reflected by the average $F_{0.5}$ -score, which, in contrast to the balanced F_1 -score, puts an emphasis on precision. In the general case, the F_β -score is calculated as

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

FlawFinder achieves the highest average $F_{0.5}$ -score in the field. A comparative overview of all participants in the competition on quality flaw prediction as determined by the organizing committee (Anderka and Stein, 2012) can be seen in table 5.9.

We carried out a detailed error analysis for each flaw in order to identify the main types of errors made by the classifier. The numbers of false positive and false negative instances according to the cross evaluation on the training set can be seen in table 5.10.

The 71 false positives for *Advert* mostly contain articles about institutions such as universities or government bodies. The descriptions of these institutions resemble the descriptions of companies. However, for companies the same way of writing is more often regarded as advert-style by Wikipedia users than for public institutions. The 94 false negatives are short articles with an average length of 690 tokens. Many of them do not exceed 250 tokens. These articles do not contain enough text to be reliably classified, since the *Advert* flaw largely relies on lexical features.

The 200 false positives for the *Notability* flaw contain many pages about individual persons, organizations, books or movies. Even Wikipedia users have difficulties to judge whether a specific subject qualifies for being included in the encyclopedia. Without world knowledge about the article topic, a reliable judgment cannot be made. Furthermore, the notability criteria in Wikipedia are highly disputed in the community and are not interpreted consistently by all users¹⁰⁰. For a large fraction of the 63 false negatives, the *Notability* template has been removed in newer revisions without a major change of the content

¹⁰⁰http://en.wikipedia.org/wiki/Deletionism_and_inclusionism_in_Wikipedia

| Flaw | System | Precision | Recall | F_1 | $F_{0.5}$ |
|-------------------|---|--|-------------|-------------|--|
| Advert | Ferschke et al. (2012b) | .853 | .826 | .839 | .847 |
| | Ferretti et al. (2012) | .736 | .929 | .821 | .768 |
| | Pistol and Iftene ^a | .047 | .582 | .086 | .058 |
| Empty section | Ferschke et al. (2012b) | .876 | .912 | .894 | .883 |
| | Ferretti et al. (2012) | .742 | .921 | .822 | .772 |
| | Pistol and Iftene ^a | .056 | 1.00 | .107 | .069 |
| No footnotes | Ferschke et al. (2012b) | .730 | .902 | .807 | .759 |
| | Ferretti et al. (2012) | .720 | .969 | .826 | .759 |
| | Pistol and Iftene ^a | .035 | .170 | .057 | .042 |
| Notability | Ferschke et al. (2012b) | .661 | .852 | .745 | .692 |
| | Ferretti et al. (2012) | .740 | .858 | .794 | .761 |
| | Pistol and Iftene ^a | .055 | .477 | .099 | .067 |
| Original research | Ferschke et al. (2012b) | .740 | .767 | .753 | .745 |
| | Ferretti et al. (2012) | .647 | .931 | .764 | .689 |
| | Pistol and Iftene ^a | .023 | .542 | .044 | .028 |
| Orphan | Ferschke et al. (2012b) | .863 | .925 | .893 | .875 |
| | Ferretti et al. (2012) | .830 | .979 | .899 | .856 |
| | Pistol and Iftene ^a | .017 | .241 | .031 | .021 |
| Primary sources | Ferschke et al. (2012b) | .736 | .866 | .796 | .759 |
| | Ferretti et al. (2012) | .717 | .923 | .807 | .751 |
| | Pistol and Iftene ^a | .052 | .423 | .093 | .063 |
| Refimprove | Ferschke et al. (2012b) | .615 | .751 | .676 | .638 |
| | Ferretti et al. (2012) | .735 | .970 | .836 | .772 |
| | Pistol and Iftene ^a | .035 | .357 | .064 | .043 |
| Unreferenced | Ferschke et al. (2012b) | .780 | .884 | .829 | .799 |
| | Ferretti et al. (2012) | .745 | .954 | .836 | .779 |
| | Pistol and Iftene ^a | .056 | 1.00 | .107 | .069 |
| Wikify | Ferschke et al. (2012b) | .678 | .844 | .752 | .706 |
| | Ferretti et al. (2012) | .742 | .737 | .740 | .741 |
| | Pistol and Iftene ^a | .056 | 1.00 | .107 | .069 |
| Average | Ferschke et al. (2012b) | .753 | .853 | .798 | .770 |
| | Ferretti et al. (2012) | .735 | .917 | .815 | .765 |
| | Pistol and Iftene ^a | .043 | .579 | .079 | .053 |

^a There is no full description available for this rule-based classification system.

It is described in short by [Anderka and Stein \(2012\)](#).

Table 5.9: Comparison of classifier performance on CLEF in terms of precision, recall and f-measure across all three participants in the competition on quality flaw detection. In addition to the balanced F_1 -score, we also report the $F_{0.5}$ -score, which puts more emphasis on precision than recall.

| | Advert | Empty Section | Notability | Original Research | Refimprove | Unreferenced | Orphan | Wikify | No Footnotes | Primary Sources |
|-----------------|--------|---------------|------------|-------------------|------------|--------------|--------|--------|--------------|-----------------|
| false positives | 71 | 73 | 200 | 93 | 71 | 158 | 71 | 250 | 164 | 164 |
| false negatives | 94 | 27 | 63 | 143 | 35 | 51 | 35 | 71 | 74 | 68 |

Table 5.10: Overview of classification errors per flaw on CLEF.

(for example in the article on the Bigfoot Trail¹⁰¹ or the Hong Kong Gold Coast¹⁰²). This suggests that the template has been incorrectly assigned to the training article by the Wikipedia users.

Many of the 158 false positives for the flaw *Unreferenced* did actually have no references at all or just contained an external links section. This suggests that the classifier correctly identified the problem, but the templates were missing in the article. The 51 false negatives are subject to the same problem. In this case, the *Unreferenced* template has been used for marking articles that suffer from the *Refimprove* flaw. For example, in the corpus version of the article “Robert Hartmann”, the used template was *Unreferenced* but it has been changed to the correct *Refimprove* template in a later version¹⁰³. Similar confusion can be observed in the misclassified instances of the other flaws related to references and citations, such as *Original Research*, *No Footnotes*, and *Primary Sources*. This suggests that the templates should be better defined and consolidated into fewer categories. Other false negative instances for *Unreferenced* are due to the inline usage of the templates that we have already discussed in section 5.2.1.1. According to the flaw definition, the template applies to articles that do not have any references. However, when used inline in the form `{{Unreferenced|section}}`, it only refers to the section it appears in, while the rest of the article may cite references¹⁰⁴. In order to account for this, each section has to be classified separately instead of the article as a whole.

According to the instructions provided by the organizers of the competition on flaw prediction, the *Orphan* template is to be assigned to any article that “has fewer than three incoming links”. Therefore, we introduced the feature `inlinks < 3`, which proved to be the most discriminative one for this flaw. However, the template description in Wikipedia states to “only place the `{{orphan}}` tag if the article has ZERO incoming links from other articles”¹⁰⁵. This discrepancy accounts for most of the false negatives, which have one or

¹⁰¹<http://en.wikipedia.org/wiki/index.php?diff=502614831&oldid=407680228>

¹⁰²<http://en.wikipedia.org/wiki/index.php?diff=502889724&oldid=461337252>

¹⁰³<http://en.wikipedia.org/wiki/index.php?diff=474150162&oldid=466987161>

¹⁰⁴<http://en.wikipedia.org/wiki/index.php?oldid=463206537>

¹⁰⁵<http://en.wikipedia.org/wiki/index.php?oldid=593368301#Criteria>

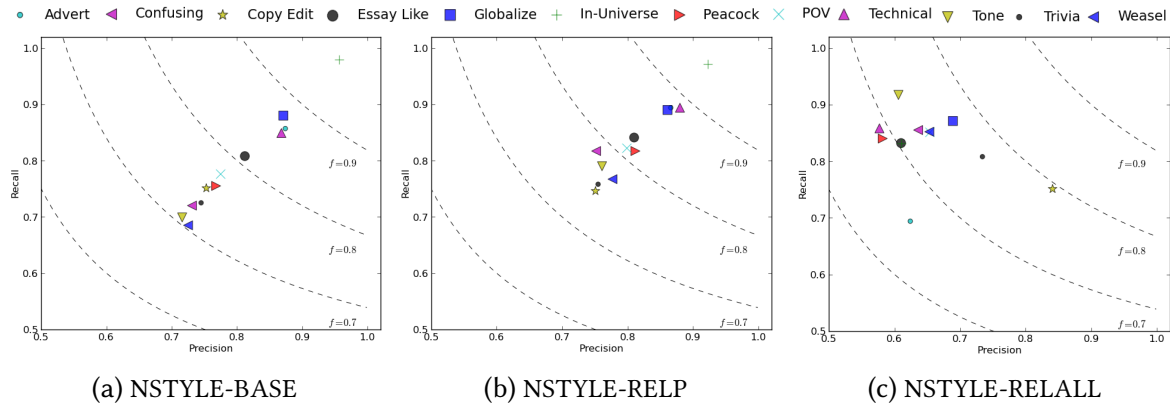


Figure 5.10: Classifier performance on NSTYLE in terms of precision, recall and F_1 -score evaluated on 10-fold cross validation with 2000 articles per flaw.

two incoming links from other articles. Removing the above mentioned feature and using the inlink counts alone can solve this issue.

The false positives for the flaw *Wikify* mainly consist of short articles. Wikification is not an issue commonly addressed in short articles, and it becomes more important as the article grows. The network and surface features used by the classifier consequently do not work well with short articles.

No regularities could be found for the misclassification of the flaw *Empty section*. It is likely that the main reason for misclassification are parsing errors. We found that sections containing mainly structured elements such as tables, infoboxes or expanded templates are particularly hard to cope with.

5.5.2.2 NSTYLE

The SVMs achieve a similar cross-validated performance on all feature sets that contain ngrams. They only showed minor improvements for individual flaws when adding non-lexical features. This suggests that the classifiers largely depend on the ngrams and that other features do not contribute significantly to the classification performance.

While structural quality flaws can be well captured by special purpose features or intentional modeling, as related work has shown, more subtle content flaws such as the neutrality and style flaws are mainly captured by the wording itself. Textual features beyond the ngram level, such as syntactic and semantic properties of the text, could further improve the classification performance of these flaws and should be addressed in future work.

Table 5.11 shows the performance of the SVMs with RBF kernel¹⁰⁶ on each dataset using the NSTYLE-NGRAM feature set. The average performance based on NSTYLE-NOWIKI is slightly lower, while using NSTYLE-ALL features results in slightly higher average F_1 -scores.

¹⁰⁶ $\gamma=0.01$ for NSTYLE-BASE and NSTYLE-RELP, $\gamma=0.001$ for NSTYLE-RELALL

| Flaw | BASE | | REL P | | REL ALL | |
|-------------|-------|-----------|-------|-----------|---------|-----------|
| | F_1 | $F_{0.5}$ | F_1 | $F_{0.5}$ | F_1 | $F_{0.5}$ |
| Advert | .866 | .871 | .880 | .872 | .657 | .637 |
| Confusing | .734 | .743 | .783 | .764 | .729 | .670 |
| Copy edit | .753 | .753 | .749 | .750 | .793 | .821 |
| Essay-like | .812 | .814 | .825 | .816 | .704 | .645 |
| Globalize | .871 | .865 | .875 | .866 | .769 | .719 |
| In-universe | .957 | .945 | .946 | .932 | .704 | .645 |
| Peacock | .768 | .776 | .815 | .813 | .687 | .620 |
| POV | .775 | .774 | .810 | .803 | .740 | .686 |
| Technical | .868 | .879 | .887 | .883 | .690 | .617 |
| Tone | .716 | .727 | .775 | .767 | .730 | .651 |
| Trivia | .745 | .758 | .756 | .756 | .769 | .748 |
| Weasel | .725 | .751 | .772 | .775 | .739 | .685 |
| \emptyset | .799 | .805 | .823 | .816 | .726 | .679 |

Table 5.11: F_β scores for the 10-fold cross validation of the SVMs with RBF kernel on all datasets using NSTYLE-NGRAM features

However, the differences are not statistically significant and thus omitted. Classifiers using the NSTYLE-NONGRAM feature set achieved average F_1 -scores below 0.50 on all datasets. The results have been obtained by 10-fold cross validation on 2,000 documents per flaw.

The classifiers trained on reliable positives and random untagged articles (NSTYLE-REL P) outperform the respective classifiers based on the NSTYLE-BASE dataset for most flaws. This confirms our original hypothesis that using the appropriate revision of each tagged article is superior to using the latest available version from the dump. The performance on the NSTYLE-REL ALL dataset, in which the topic bias has been factored out, yields lower F_1 -scores than the two other configurations. Flaws that are restricted to a very narrow set of topics (i.e. A_{topic} in figure 5.5b is small), such as the *in-universe* flaw, show the biggest drop in performance. Since the topic bias plays a major role in the quality flaw detection task, as we have shown earlier, the topic-controlled classifier cannot take advantage of the topic information, while the classifiers trained on the other corpora can make use of these characteristics as the most discriminative features. In the NSTYLE-REL ALL setting, however, the differences between the positive and negative instances are largely determined by the flaws alone. Classifiers trained on such a dataset therefore come closer to recognizing the actual quality flaws, which makes them more useful in a practical setting despite lower cross-validated scores.

In addition to cross-validation, we performed a cross-corpus evaluation of the classifiers for each flaw. Therefore, we evaluated the performance of the unbiased classifiers (trained

on RELALL) on the biased data (NSTYLE-RELP) and vice versa. Hereby, the positive training and test instances remain the same in both settings, while the unbiased data contains negative instances sampled from A_{rel} and the unbiased data from A_{rnd} (see figure 5.5). With the NSTYLE-NGRAM feature set, the reliable classifiers outperformed the unreliable classifiers on all flaws that can be well identified with lexical cues, such as *Advert* or *Technical*. In the biased case, we found both topic related and flaw specific ngrams among the most highly ranked ngram features. For example, in the case of the flaw *Technical*, we saw many general ngrams related to mathematics and science and technical terms from these areas. In the unbiased case, most of the informative ngrams were flaw specific. In the example of *Technical* articles, we mainly observed technical terms. Consequently, biased classifiers fail on the unbiased dataset in which the positive and negative classes are sampled from the same topics, which renders the highly ranked topic ngrams unusable. Flaws that do not largely rely on lexical cues, however, cannot be predicted more reliably with the unbiased classifier. This means that additional features are needed to capture these flaws. We tested this hypothesis by using the full feature set NSTYLE-ALL and saw a substantial improvement on the side of the unbiased classifier because of the added features, while the performance of the biased classifier remained unchanged. This indicates that the predictive power of the biased classifier mainly depends on the generic ngram features, which capture the topic cues in the dataset, while it cannot be improved with the additional features. Since the topic bias is ruled out in the unbiased case, the generic ngrams are less efficient and the classifier can gain from the additional features.

A direct comparison of our results to related work is difficult, since neutrality and style flaws have not been targeted before in a similar manner. However, the *Advert* flaw was also part of the ten flaw types in the PAN Quality Flaw Recognition Task (Anderka and Stein, 2012). The best system achieved an F_1 score of 0.839, which is just below the results of our system on the NSTYLE-BASE dataset, similar to the PAN setup.

5.6 Mining Flaw Corrections from the Revision History

In addition to finding articles with potential quality flaws, another important use case of automatic quality flaw recognition is the identification of the quality problems at a specific position within a given article. We therefore transfer the quality flaw recognition task from the article level to the sentence level.

In order to build a sentence classifier for a given flaw type, we have to create corpora of quality flaw corrections, i.e. pairs of flawed and flawless sentences. Analogously to the experiments on the article level, the training instances have to be reliable. We therefore follow a similar approach as in the previous experiment and extract pairs of flawed and flawless article revisions from the revision history as described in section 5.3. Instead of using article-scope templates in the corpus creation process, we now use inline- and

| # | Flawed | Revised |
|----|--|--|
| S1 | Some believe that Iran controls the majority of terrorism in Israel. | Israeli intelligence believes that Iran controls the majority of terrorism in Israel. |
| S2 | The adaptation was highly praised and was subsequently released on audio cassette. | The adaptation was subsequently released on audio cassette. |
| S3 | The group were told that the Ghost would not come as they were making too much noise. | Moore told the group the ghost would not come as they were making too much noise. |
| S4 | The Iraq War troop surge of 2007 was part of this "new way forward" and has been credited by some with a dramatic decrease in violence and an increase in political and communal reconciliation in Iraq. | The Iraq War troop surge of 2007 was part of this "new way forward". |
| S5 | According to theorists, there are many signs that will confirm these claims. | According to theorists, such as David Icke, there are many signs that will confirm these claims. |
| S6 | In one of his major works he also showed that Indian philosophy, once translated into standard academic language, is worthy of being called philosophy by Western standards. | He wrote books on Indian philosophy according to Western academic standards, and made Indian philosophy worthy of serious consideration in the West. |

Table 5.12: Sample sentence pairs from the uncertainty corpus

section-scope templates (see section 5.2.1.1), which mark specific sentences or sections as flawed. We align each pair of flawed and flawless revisions on a section level by matching the section headlines in both revisions and using *Greedy String Tiling* (Wise, 1996) for comparing the section texts. We then discard all section pairs that do not contain any flaw templates. The remaining section pairs are split into sentences and processed with a text difference (*diff*) algorithm (Myers, 1986). From the so aligned sentences, we finally extract all pairs that contain at least one flaw template in order to create a parallel corpus of flawed sentences and their corrections.

A similar approach has been suggested by Recasens et al. (2013), who extract edits from the article revision history which meant to remove bias from the article text. To this end, the authors retrieve all articles that are or, at any point in time had been, members of the Wikipedia NPOV category indicating that their neutrality is disputed. From the revision histories of these articles, they extract all commits that contained a comment mentioning POV or NPOV, thus hinting at the neutrality dispute. They finally discard all edits that merely added a URL or changed less than four characters.

We carried out experiments for the flaws Weasel and Peacock, which we combined into a single corpus due to their similar purpose which is concerned with textual uncertainty. The extracted uncertainty corpus contains 16,241 sentence pairs. Table 5.12 shows a set of sample sentence pairs. Upon manual inspection of a random sample of 200 sentence pairs, we identified six main types of corrections for uncertainty flaws: pronoun replacement (S1), intensifier removal (S2), passive-active transformation (S3), clipping (S4), expansion (S5), paraphrasing (S6).

On this corpus, we trained a binary uncertainty classifier for sentences using the Flaw-Finder system described in section 5.4 using only NGRAM features. With this baseline approach, we achieved an F_1 -Score of 0.65 in identifying sentences expressing uncertainty. The task carried out in this experiment is similar to the uncertainty detection subtask

(Task1W) of the CONLL 2010 shared task on hedge detection (Farkas et al., 2010). The winning team of that task achieved a performance of $F_1 = 0.60$ on identifying uncertain sentences using a dictionary of hedge cues as their most predictive features. On the same corpus, FlawFinder achieved an F_1 -Score of 0.59 using the same configuration as on the parallel corpus described above. Even though the two corpora are not directly comparable, they are similar enough to allow the tentative conclusion that the parallel data mined with our approach helps to improve the classification performance compared to a corpus that does not contain explicit negative instances. This goes against the intuition we have gained in the document-level experiments, where the cross-validated performance dropped on the unbiased dataset. The main reason for this is that the shorter texts in the sentence corpora are less affected by the topic bias than longer documents, which was the main reason for the unrealistically high cross-validated performance.

Apart from the sentence-level classification task, the proposed approach for mining flawed sentences and their corrections from Wikipedia gives rise to many opportunities in analytical linguistic research. Such corpora can easily be created for any inline- or section-scope cleanup templates, which makes it possible to obtain parallel corpora for a wide range of linguistic phenomena.

5.7 Limitations in the Predictability of Quality Flaws

In section 5.2.3, we have evaluated how well a human annotator can manually perform the quality flaw prediction task in order to gain an impression of the reliability of cleanup templates as quality flaw markers. The study showed that flaws which usually affect only parts of an article are harder to detect by humans than flaws that affect the article as a whole. This was not necessarily the case in the machine learning experiments. We now return to this issue and review the limitations of cleanup templates as the basis for training quality flaw classifiers on a broader level.

We have already established in chapter 4 that quality dimensions have to be measurable in order to be useful for information quality management. In particular, we discussed the *consistency*, *subjectivity*, *operationalizability* and *interpretability* of quality dimensions. Three of these properties can also be directly translated to the quality flaw prediction task and can be described by the following questions:

Flaw Subjectivity: *How much room for interpretation does a given template or flaw definition allow and how consistent are the label assignments across different raters?*

Flaw Operationalizability: *Is it possible to identify descriptive features that are predictive of the given quality flaw and that can be automatically extracted from the available data, i.e. the article and its metadata?*

Flaw Consistency: *Given that a set of features has been identified that are indicative of a given flaw, do the same feature values indicate the presence or absence of the flaw consistently in every possible situation?*

In order to illustrate the meaning of these three properties, we discuss two exemplary quality flaws from the CLEF and NSTYLE corpora in the light of their subjectivity, operationalizability and consistency.

The *Orphan* flaw identifies articles that are not connected to other articles via hyperlinks. Given only this basic definition, the decision whether an article should or should not be marked with the flaw is purely objective without any room for interpretation or personal judgment. In practice, however, the orphan criteria¹⁰⁷ include cases of weakly linked articles or small cliques of articles and leave the decision whether to assign the label in these fringe cases up to the community. Hence, the orphan flaw cannot be considered purely objective but is subjective to a low degree. Since the properties that characterize an article as orphaned are all governed by technical features such as incoming and outgoing links, the flaw has a high degree of operationalizability. The consistency of the flaw is only influenced by the fringe cases included in its definition, i.e. whether or not a weakly linked article is considered orphaned. Apart from that, the clear-cut features result in a high degree of consistency.

The *Technical* flaw identifies articles that are written in an overly technical tone that prevents the general audience from understanding it. The subjectivity of its definition already becomes clear in the message box of this flaw which states

*This article may be too technical for most readers to understand. Please help improve this page to make it understandable to non-experts, without removing the technical details. The talk page may contain suggestions.*¹⁰⁸

In fact, whether or not an article can be considered to be too technical for a general audience heavily depends on the familiarity of the article maintainer with the subject matter and on their perception of the level of understanding that the general audience possesses. Even more severe is the issue of flaw consistency in the case of *Technical* articles. Since the notion of this flaw is not absolute but relative to the article topic, i.e. the article text is considered to be more technical than it needs to be, we need to consider the article topic as a frame of reference. For example, it might not be possible to write a comprehensive article about a technical subject without ample use of technical vocabulary, while the same number of technical terms might be considered excessive in an equally long non-technical article. Moreover, terms that might be considered overly technical in one article might be necessary vocabulary in another and considered appropriate in that context. Incorporating the topic

¹⁰⁷<http://en.wikipedia.org/wiki/WP:0>

¹⁰⁸<http://en.wikipedia.org/wiki/index.php?oldid=582079420>

as a frame of reference is therefore important when modeling flaws that are interpreted differently across different subject areas in Wikipedia.

5.8 Chapter Summary

In this chapter, we presented an approach to automatically identify quality flaws in Wikipedia articles by means of cleanup template prediction. While cleanup templates are good proxies for quality flaws and thus a viable resource for compiling quality flaw corpora as training data for machine learning classifiers, we found that many templates exhibit a topic bias that negatively influences the classifier performance and even biases manual analyses.

We found that certain templates exhibit a topical preference, i.e. they tend to occur in articles about particular topics, or even show a topical restriction, i.e. the templates exclusively occur in articles about particular topics. This fact has to be taken into account when sampling the data for quality flaw corpora in order to avoid a topic bias that influences both any data analyses and machine learning classifiers trained on this data.

We therefore introduced an approach to extract reliable positive and negative training instances from the article revision history which factors out the topic bias and improves the overall data quality.

We furthermore presented a corpus of articles with neutrality and style flaws that has been sampled with this technique. Our machine learning experiments on this corpus show that the reliable classifiers tend to exhibit a lower cross-validated performance than classifiers trained on the biased datasets but the scores more closely resemble their actual performance in practical settings.

Finally, we described an approach for mining quality flaw corrections from the revision history. This method can both be used to create a new parallel corpus of flawed and flawless language as well as for identifying the position of quality flaws within articles rather than merely identifying flawed articles.

We closed the chapter by discussing the limitations of the quality flaw prediction task based on cleanup templates. While some flaws are predictable on a global scale using all available training data, like most structural and organizational flaws, others have to be considered within a narrow context of the subject area they are used in. That is, the features which indicate a flaw in one subject area might not be predictive of the same flaw in another subject area and rather result in a higher rate of false positives.

CHAPTER 6

Dialog Analysis of Wikipedia Talk Pages

“A conversation is a dialogue, not a monologue. That’s why there are so few good conversations: due to scarcity, two intelligent talkers seldom meet.”

— Truman Capote

Every article in Wikipedia has an associated discussion page – or Talk page – on which the active contributors discuss the future development of the article, coordinate their work and collaboratively decide how conflicting plans regarding the improvement of the article should be resolved. In this chapter, we discuss how the information on the article Talk pages can be leveraged for information quality management purposes and how an analysis of these pages provides us with insights into the collaborative writing process that complements the knowledge we can gain from analyzing the article revision history.

We first give an overview of our motivation (section 6.1) and present the theoretical background for our work (section 6.2) We then discuss related work on computational dialog analysis in general and the analysis of Wikipedia discussions in particular (section 6.3). We proceed with a detailed examination of two Wikipedia discussion corpora (section 6.4) which we annotated with an annotation scheme to capture the coordination efforts for article improvement (section 6.5). Finally, we investigate how these corpora can be used to automatically tag unseen discussions with dialog act labels which identify quality problems discussed on the Talk pages and the solutions proposed by the contributors (section 6.6). We conclude the chapter with an overview of a real world application of the Talk page classification system (section 6.7) and a summary of our findings (section 6.8).

6.1 Motivation and Overview

The article Talk pages in Wikipedia are the main communication hub for the discussions related to article improvement, work coordination and quality assessment. On these pages, decisions are made that shape the evolution of the associated articles and have a vital impact on their quality. As we have discussed in chapter 3.5, Talk pages are largely unstructured wiki pages which mimic the appearance of traditional threaded web forums. The disparity between the lack of explicit structure on the one hand and the structured form it seeks to resemble on the other hand is one of the main reasons why these Talk pages are difficult to use by novice users (Schneider et al., 2011) and why they are hard to process computationally (Ferschke et al., 2012a).

As studies have shown, the overall discussion activity in Wikipedia is on the rise, indicated by an increasing proportion of Talk page edits, while the relative number of article edits constantly declines (Schneider et al., 2010; Stvilia et al., 2008). High discussion activity naturally results in fast growing Talk pages. However, threads that reach a certain size and age and that are inactive for some time are automatically archived and thus no longer directly visible on the main article Talk page. While the archived information is technically retained, the rudimentary search capabilities¹⁰⁹ and the lack dialog structure do not allow users to easily retrieve old content from the archives, which effectively renders old information lost.

As a consequence, important decisions once made about an article are likely to be forgotten over time and many recurring issues and topics have to be discussed over and over again which unnecessarily binds much of the available workforce. Furthermore, while Talk pages are directly connected to individual articles, topics discussed in the context of one article might still be relevant for related articles. However, there is no easy way to make this connection as a user without actively monitoring the Talk pages of all related articles.

Finally, even though Talk pages are mainly used by Wikipedia users who actively contribute to the encyclopedia, these discussions often hold information that could be interesting to the general public. Linking the information on the Talk pages to the related sections in the corresponding article could help casual users of Wikipedia to gain access to this additional information source. However, without a basic understanding of the discourse structure of the Talk pages, it is not possible to establish this link automatically or semi-automatically.

Extracting the essential information about work coordination with a particular focus on the article quality improvement activities will help to gain an overview of the relevant topics covered and the decisions made in past discussions and thus improve the consistency

¹⁰⁹By default, no search in Talk pages archives is available. A rudimentary search function can be manually enabled by including the corresponding template (e.g. *Template:Search_archives*) on the main Talk page of an article.

and availability of this important information across Wikipedia. Such a system will take a key role in the global quality management process and even holds opportunities to improve the user experience of the encyclopedia. One of the main challenges on the way to achieve this goal is to overcome the unstructured nature of the Talk pages and to reliably segment the dialog while retaining relevant meta information, such as the identity of the contributors, the time stamp of each contribution and the basic thread structure. This is the prerequisite for employing semi-supervised machine learning to classify the contributions into predefined categories tailored towards quality assessment activities, which is the main topic of this chapter.

From a scientific point of view, article Talk pages are a unique type of web discourse and a valuable resource for the humanities and writing sciences, since the discussions develop in parallel with the discussed articles and provide insights into the meta level of the collaborative writing process that normally remains hidden. With structured access to this resource, the linguists and researchers in the writing sciences have the unparalleled possibility to observe these hidden processes without having to conduct interviews or carrying out supervised field experiments.

The main contributions of this chapter can be summarized as follows:

Contribution 6.1: *We present an algorithm for dialog segmentation of Wikipedia article discussions based on the revision history (section 6.4.1).*

Contribution 6.2: *We compile two corpora of Wikipedia article discussions from the Simple English Wikipedia and the English Wikipedia (section 6.4).*

Contribution 6.3: *We introduce annotation schemes for annotating turns in article discussions to capture the coordination efforts of article improvement (section 6.5).*

Contribution 6.4: *We annotate the corpora with our newly developed annotation schemes and analyze the resulting datasets (section 6.5).*

Contribution 6.5: *We develop a system for automatically labeling turns in Wikipedia article discussions with dialog act labels from our annotation scheme and evaluate the performance of the classifiers (section 6.6).*

6.2 Linguistic Background

Early models of human communication predominantly followed a positivist view, which regards logic and reason to be the governing principle of language. It postulates that human language merely serves as a passive container of meaning that can readily be extracted by anyone able to decode the signs, i.e. the words of the language (Krippendorff, 1994).

This rather simplistic view of language and communication was soon superseded by more advanced models following the theoretical paradigm of structural linguistics, which acknowledge that language serves different functions at the same time. Human communi-

cation has therefore to be analyzed on multiple levels simultaneously under consideration of the context and participants in the communication (Bühler, 1934; Jakobson, 1960).

Speech Act Theory. John Austin finally shifted the focus of linguistics from the mere declarative use of language as a means for making factual statements towards its non-declarative use as a tool for performing actions. In his influential work “How to do things with words” (1962), Austin argues that a large part of human language goes beyond mere statements, assertions, or propositions and involves the performance of actions. He furthermore claimed that there are utterances that cannot be analyzed in terms of truth conditions, on which most previous theories have relied. This newfound concept of language as a means to perform actions ultimately led to the theory of speech acts. Austin defines that language performs actions on three levels simultaneously, the *locutionary level*, the *illocutionary level* and the *perlocutionary level*.

The locutionary act describes the performance of the utterance itself, i.e. its verbalization and pronunciation. At the same time, the illocutionary act concerns the pragmatic level of the utterance, i.e. it captures the intention of the speaker. Finally, the perlocutionary act is directed at the recipient of the message and the effect the utterance has achieved on him. Only by analyzing communication on all three levels it is possible to achieve a full understanding of its meaning. Austin’s speech act theory was further systematized by Searle (1969), who, among other refinements of the theory, introduced a taxonomy of illocutionary acts (Searle, 1976). He distinguishes between five different illocutionary classes

Assertives/Representatives: *communicate a proposition which the sender of the message believes to be true*

Example: *It’s raining today.*

Directives: *cause the recipient to perform an action*

Example: *Close the door!*

Commissives: *commit the sender to perform an action in the future*

Example: *I’ll be back.*

Expressives: *expresses the sender’s attitude or emotions towards the proposition of the utterance*

Example: *Good job, congratulations.*

Declarations: *constitute an act that directly changes reality*

Example: *I now pronounce you husband and wife.*

This taxonomy of illocutionary acts has become an important instrument for the analysis of human utterances and has often been used as the starting point for the development of new schemes for speech act analyses, both in traditional linguistic (Sadock, 2006) and in computational linguistics (Jurafsky, 2006).

Dialog Acts. According to [Bunt and Black \(2000\)](#), classifying utterances with respect to the performed speech acts promises deep insights into the pragmatic structures of the discourse. The concept is of particular importance for the analysis of *human-human dialog*. While a *monologue* can be considered a form of unidirectional communication between a sender (writer, speaker) and a receiver (reader, listener), we define a *dialog* as the bidirectional communication between multiple agents who exchange coherent messages and switch between the roles of sender and receiver in the course of the communication. Although it is possible to have multiple senders at the same time (e.g. several people talking or writing at the same time), we only consider the case of a single sender at the same time. As long as an agent is assigned the role of the sender, it is considered to be his or her *turn*. Passing the sender role to each other is therefore defined as *turn-taking*.¹¹⁰

In dialog settings, speech acts are usually referred to as *dialog acts*. The exact definition of this term differs across the literature ([Bunt and Black, 2000](#); [Jurafsky, 2006](#)), but can be summarized as *a specialized speech act defining the function of an utterance in the context of a particular dialog*. Other terms, such as *communicative acts*, *conversation acts*, *conversational moves* or *dialog moves* roughly translate to similar concepts ([Traum, 2000](#)).

Dialog Act Identification for IQ Management. Since the discussions on Wikipedia article Talk pages mainly revolve around the development of the associated articles and the improvement of their quality, identifying dialog acts tailored towards this kind of discourse can help to identify and organize the main intentions of the contributions in these discussions. Instead of applying the generic classification scheme of illocutionary speech acts proposed by [Searle](#), we have to define a more fine grained set of specialized dialog acts that satisfy the requirements of the information quality management setting. We propose such a scheme in section 6.5 of this chapter.

6.3 Related Work

While the linguistic theory provides the theoretical framework for computationally analyzing human dialog, it is necessary to operationalize the linguistic concepts in a concrete scheme in order to annotate, process, and analyze real-world examples of human dialog.

A well known, domain- and task-independent annotation scheme is DAMSL – Dialog Act Markup in Several Layers ([Core and Allen, 1997](#)). It was created as the standard annotation scheme for dialog tagging on the utterance level by the Discourse Resource Initiative. It uses a four-dimensional tagset that allows arbitrary label combinations for each utterance. [Jurafsky et al. \(1997\)](#) augmented the DAMSL scheme to fit the peculiarities of the

¹¹⁰For the sake of brevity, a more detailed introduction to conversation analysis, the mechanics of dialog situations, conversational implicature or indirect speech acts has been omitted. A comprehensive overview can be found in [Hutchby and Wooffitt \(2008\)](#).

Switchboard corpus. The resulting SWDB-DAMSL scheme contained more than 220 distinct labels which have been clustered to 42 coarse grained labels. Both schemes have often been adapted for special purpose annotation tasks.

While DAMSL was originally designed for annotating transcripts of spoken dialog, a large part of current research is directed at written online discourse. In addition to analyzing web forums (Kim et al., 2010a), chats (Carpenter and Fujioka, 2011) and emails (Cohen et al., 2004), Wikipedia Talk pages have recently moved into the center of attention of the research community.

In the remainder of this section, we will discuss the different aspects of Wikipedia discussions that have been investigated in related work and which are highly relevant for our efforts to analyze quality management activities on Talk pages.

6.3.1 Work Coordination and Conflict Resolution

Viégas et al. (2007) were among the first to draw attention to Wikipedia Talk pages as an important resource in its own right. In an empirical study, they discovered that articles with Talk pages have, on average, 5.8 times more edits and 4.8 times more participating users than articles without any Talk activity. Furthermore, they found that the number of new Talk pages increased faster than the number of content pages. In order to better understand how the rapidly increasing number of Talk pages are used by Wikipedians, they performed a qualitative analysis of selected discussions. The authors manually annotated 25 “purposefully chosen”¹¹¹ Talk pages with a set of 11 labels in order to analyze the aim and purpose of each user contribution. Each turn was tagged with one of the following labels:

- request for editing coordination
- request for information
- reference to vandalism
- reference to Wikipedia guidelines
- reference to internal Wikipedia resources
- off-topic remark
- poll
- request for peer review
- information boxes
- images
- other

The first two categories, requests for coordination (58.8%) and information (10.2%), were most frequently found in the analyzed discussions, followed by off-topic remarks (8.5%),

¹¹¹According to Viégas et al. (2007), “[t]he sample was chosen to include a variety of controversial and non-controversial topics and span a spectrum from hard science to pop culture.”

guideline references (7.9%), and references to internal resources (5.4%). This shows that Talk pages are not used just for the “retroactive resolution of disputes”, as the authors hypothesized in their preliminary work (Viégas et al., 2004); rather, they are used for proactive coordination and planning of the editorial work.

Schneider et al. (2010, 2011) pick up on the findings of Viégas et al. and manually analyze 100 Talk pages with an extended annotation scheme. In order to obtain a representative sample for their study, they define five article categories to choose the Talk pages from: *most-edited articles*, *most-viewed articles*, *controversial articles*, *featured articles*, and a *random set of articles*. In addition to the 11 labels defined by Viégas et al., Schneider et al. classify the user contributions as

- references to sources outside Wikipedia
- references to reverts, removed material or controversial edits
- references to edits the discussant made
- requests for help with another article

The authors evaluated the annotations from each category separately and found that the most frequent labels differ between the five classes. Characteristic peaks in the class distribution could be found for the “reverts” label, which is a strong indicator for discussions of controversial articles. Interestingly, the controversial articles did not have an above-average discussion activity, which was initially expected due to a high demand of coordination. The labels “off-topic”, “info-boxes”, and “info-requests” peak in the random category, which are apt to contain shorter Talk pages than the average items from the other classes. In accordance with Viégas et al., coordination requests are the most frequent labels in all article categories, running in the 50% to 70% range. The observed distribution patterns alone are not discriminative enough for identifying the type of article a Talk page belongs to, but they nevertheless serve as valuable features for the Talk page analysis.

Furthermore, the labels can be used to filter or highlight specific contributions in a long Talk page to improve the usability of the Talk platform. Schneider et al. (2011) perform a user study in which they evaluate a system that allows discussants to manually tag their contribution with one of the labels. Most of the 11 participants in the study perceived this as a significant improvement in the usability of the Talk page, which they initially regarded as confusing. Given enough training data, this classification task can be tackled automatically using machine learning algorithms.

In a large-scale quantitative analysis, Kittur et al. (2007) confirm earlier findings by Viégas et al. and demonstrate that the amount of work on content pages in Wikipedia is decreasing while the *indirect work* is on the rise. They define indirect work as “excess work in the system that does not directly lead to new article content.” Besides the efforts for work coordination, indirect work comprises the resolution of conflicts in the growing community of Wikipedians. In order to automatically identify conflict hot spots or even

Table 6.1: Page-level features proposed by Kittur et al. (2007)

| Feature | Page |
|---|---|
| Revisions ^a | Article ⁴ , Talk ¹ , Article/Talk |
| Page length | Article, Talk, Article/Talk |
| Unique editors ^a | Article ⁵ , Talk, Article/Talk |
| Unique editors ^a /Revisions ^a | Article, Talk ³ |
| Links from other articles ^a | Article, Talk |
| Links to other articles ^a | Article, Talk |
| Anonymous edits ^{a,b} | Article ⁷ , Talk ⁶ |
| Administrator edits ^{a,b} | Article, Talk |
| Minor edits ^{a,b} | Article, Talk ² |
| Reverts ^{a,c} | Article |

^a Raw counts ¹⁻⁷ Feature utility rank
^b Percentage
^c By unique editors

to prevent future disputes, the authors developed a model of conflict on the article level and demonstrate that a machine learning algorithm can predict the amount of conflict in an article with high accuracy. In contrast to the works discussed above, Kittur et al. do not employ a hand-crafted coding scheme to generate a manually annotated corpus; rather, they extract the “*controversial*” tags that have been assigned to articles with disputed content by Wikipedia editors. This human-labeled conflict data is obtained from a full Wikipedia dump with all page revisions (*revision dump*) using the Hadoop¹¹² framework for distributed processing. The authors define a measure called *Controversial Revision Count* (CRC) as “the number of revisions in which the ‘controversial’ tag was applied to the article”. These scores are used as a proxy for the amount of conflict in a specific article and are predicted by a Support Vector Machine regression algorithm from raw data. The model is trained on all articles that are marked as controversial in their latest revision and evaluated by means of five-fold cross validation. As features, the authors define a set of page-level metrics based on both articles and talk pages (see table 6.1). They evaluated the usefulness of each feature, which is indicated by the individual ranks as superscript numbers in the table.

The authors report that the model was able to account for almost 90% of the variation in the CRC scores ($R^2 = 0.897$). They furthermore validate their model in a user study by having Wikipedia administrators evaluate the classification results on 28 manually selected articles that have not been tagged as controversial. The results of this study showed that the CRC model generalizes well to articles that have never been tagged as controversial. This opens up future applications like identifying controversial articles before a critical point is reached.

¹¹²<http://hadoop.apache.org>

6.3.2 Authority and Social Alignment

Discussions on article Talk pages often aim at keeping articles in line with Wikipedia’s guidelines for quality, neutrality and notability. The outcome of these discussions therefore can have a big impact on the future development of an article. If such a discussion is not grounded in authoritative facts but rather in subjective opinions of individual users, a dispute about content removal, for example, may lead to the unjustified removal of valuable information. Wikipedia Talk pages are, for the most part, pseudonymous discussion spaces and most of the discussants do not know each other personally. This raises the question how the users of Talk pages decide which claim or statement in a discussion can be trusted and whether an interlocutor is reliable and qualified.

[Oxley et al. \(2010\)](#) analyze how users establish credibility on Talk pages. They define six categories of *authority claims* with which users account for their reliability and trustfulness (see table 6.2). Based on this classification, [Bender et al. \(2011\)](#) created a corpus of social acts in Wikipedia Talk pages (AAWD). In addition to authority claims, the authors define a second annotation layer to capture *alignment moves*—i.e. expressions of solidarity or signs of disagreement among the discussants. At least two annotators labeled each of the 5,636 turns extracted from 47 randomly sampled Talk pages from the English Wikipedia. The authors report an overall inter-annotator agreement of $\kappa = 0.59$ for authority claims and $\kappa = 0.50$ for alignment moves.

[Marin et al. \(2011\)](#) use the AAWD corpus to perform machine learning experiments targeted at automatically detecting authority claims of the *forum* type (cf. Table 6.2) in unseen discussions. They particularly focus on exploring strategies for extracting lexical features from sparse data. Instead of relying only on n -gram features, which are prone to overfitting when used with sparse data, they employ knowledge-assisted methods to extract meaningful lexical features. They extract word lists from Wikipedia policy pages to capture policy-related vocabulary and from the articles associated with the Talk pages to capture vocabulary related to editor discussions. Furthermore, they manually create six word lists related to the labels in the annotation scheme. Finally, they augment their features with syntactic context gained from parse trees in order to incorporate a higher level linguistic context and to avoid the explosion of the lexical feature space that is often a side effect of higher level n -grams. Based on these features, the authors train a maximum entropy classifier to decide for each sentence whether it contains a forum claim or not.¹¹³ The decision is then propagated to the turn level if the turn contains at least one forum claim. The authors report an F_1 -score for the evaluation set of 0.66. Besides being a potential resource for social studies and online communication research, the AAWD corpus and approaches to

¹¹³The corpus was split into training set (67%), development set (17%) and test set (16%).

| Claim type | Based on |
|----------------------------|--|
| Credentials | Education, Work experience |
| Experiential | Personal involvement in an event |
| Institutional ^a | Position within the organizational structure |
| Forum | Policies, Norms, Rules of behavior (in Wikipedia) |
| External | Outside authority or resource |
| Social Expectations | Beliefs, Intentions, Expectations of social groups |

^a Not encoded in the AAWD corpus

Table 6.2: Authority claims proposed by [Oxley et al. \(2010\)](#) and [Bender et al. \(2011\)](#)

automatic classification of social acts can be used to identify controversial discussions and online trolls.¹¹⁴

In an attempt to investigate how users of different status groups interact, [Danescu-Niculescu-Mizil et al. \(2012\)](#) created a corpus of 5,657 Talk pages with overall 125,292 discussion threads. Their hypothesis was that the amount of language coordination in a conversation will depend on the social status of the participants. The authors define language coordination as the stylistic mimicry of the interlocutor which describes the tendency of a person to adapt the usage of function words of his or her communication partner. Social status in the Wikipedia context is furthermore defined as the user role of the discussants (see chapter 3.3.1), such as *registered user* or *admin*. The authors find that people with a lower social status exhibit a greater tendency to language coordination than users with more power and that a change in status will also trigger a change in the coordination behavior. Furthermore, the intention to convince a communication partner with an opposing view of their own opinion will result in a power deficit and thus trigger a higher level of language coordination. This effect can not only be observed in Wikipedia, but is a stable phenomenon in other kinds of communication such as Supreme Court meetings ([Danescu-Niculescu-Mizil et al., 2012](#)).

6.3.3 User Interaction

It is not only the content of Talk pages which has been the focus of recent research, but also the social network of the users who participate in the discussions. [Laniado et al. \(2011\)](#) create Wikipedia discussion networks from Talk pages in order to capture structural patterns of interaction. They extract the thread structure from all article and user Talk pages in the English Wikipedia and create tree structures of the discussions. For this, they rely on user signatures and turn indentation. The authors consider only registered users, since IP ad-

¹¹⁴A *troll* is a participant in online discussions with the primary goal of posting disruptive, off-topic messages or provoking emotional responses.

addresses are not unique identifiers for the discussants. In the directed article reply graph, a user node A is connected to a node B if A has ever written a reply to any contribution from B on any article Talk page. They furthermore create two graphs based on User Talk pages which cover the interactions in the personal discussion spaces in a similar manner.

The authors analyze the directed degree assortativity of the extracted graphs. In the article discussion network, they found that users who reply to many different users tend to interact mostly with inexperienced Wikipedians while users who receive messages from many users tend to interact mainly with each other. They furthermore analyzed the discussion trees for each individual article, which revealed characteristic patterns for individual semantic fields. This suggests that tree representations of discussions are a good basis for metrics characterizing different types of Talk pages, while the analysis of User Talk pages might be a good foundation for identifying social roles by comparing the different discussion fingerprints of the users.

A different aspect of the social network analysis in Wikipedia is examined by [Massa \(2011\)](#). He aims at reliably extracting social networks from User Talk pages. Similarly to [Laniado et al. \(2011\)](#), he creates a directed graph of user interactions. The interaction strength between two users is furthermore quantified by weighted edges with weights derived from the number of messages exchanged by the users. The study is based on networks extracted from the Venetian Wikipedia. [Massa](#) employ two approaches to extract the graphs automatically, one based on parsing user signatures (signature-based approach) and the other one based on the revision history regarding every commit by a user as an individual message (history-based approach). He compares the results with a manually created gold standard and finds that the revision based approach produces more reliable results than the signature-based approach, which suffers from the extreme variability of the signatures. However, history-based processing often resulted in higher weights of the edges, because several edits of a contribution are counted as individual messages. [Massa](#) furthermore identifies several factors that impede the network extraction, like noise in the form of bot messages and vandalism, inconsistently used usernames, and unsigned messages. While these insights might be a good basis for future work on network extraction tasks, they are limited by the small Venetian Wikipedia on which the study is based. Talk pages in larger Wikipedias are much longer, more complex and are apt to contain pitfalls not recognized by this work.

6.3.4 Information Quality

Related work that uses Wikipedia Talk pages for information quality analyses in Wikipedia is scarce, but most relevant for the work presented in this thesis. As we have argued before, the information on Talk pages contains valuable insights into the readers' judgments of articles and comments about their potential deficiencies.

Stvilia et al. (2005, 2007, 2008) analyze 60 discussion pages in order to identify which types of information quality (IQ) problems have been discussed by the community. Based on this analysis, they determine twelve IQ problems along with a set of related causal factors for each problem and actions that have been suggested by the community to tackle them. For instance, IQ problems in the quality dimension *complexity* may be caused by low readability or complex language and might be tackled by replacing, rewriting, simplifying, moving, or summarizing the problematic article content. They furthermore identify trade-offs among these quality dimensions of which the discussants on Talk pages are largely aware. For example, an improvement in the dimension *completeness* might result in a deterioration in *relevance*, i.e. the more details are added to an article, the higher is the chance to incorporate irrelevant information.

To the best of our knowledge, no previous work has yet attempted to use machine learning to automatically classify user contributions in Wikipedia Talk pages with respect to the article improvement efforts they express. This is the subject of our work presented in this chapter.

6.4 Wikipedia Article Discussion Corpora

For our experiments, we created two corpora of Wikipedia Talk pages from the Simple English Wikipedia (SEWD corpus) and the English Wikipedia (EWD corpus). While the English Wikipedia is the largest language version (see figure 3.1) with the biggest community, the Simple English Wikipedia is a small, special purpose Wikipedia that hosts articles written in a basic English vocabulary and a simple syntax so they are easy to understand by non-native speakers. In the remainder of this section, we first describe the data extraction and discourse segmentation techniques used for retrieving and preprocessing the data for the corpora. We then discuss the sampling strategies taken for selecting the documents. In the subsequent section (6.5), we introduce the annotation schemes used for annotating the corpora, describe the annotation process and provide a detailed corpus analysis.

As already discussed in chapter 3.6, there are multiple possibilities to access Wikipedia programmatically. The best approach depends on the demands of the particular task at hand. Similarly to the experiments in chapter 5, we take a hybrid approach in which we combine a raw text corpus containing the desired Talk pages with a preprocessed database of the full Wikipedia from which the pages have been extracted. This way, while having a fixed corpus for human annotation, we can use the corresponding Wikipedia database at any time to directly retrieve additional information about the Talk pages, the associated articles or the involved users (see figure 6.1). We use the Java Wikipedia Library (JWPL) for creating the Wikipedia database representations from the freely available Wikipedia XML data dumps paired with the Wikipedia Revision Toolkit (WRT) to include the revision history of all articles and Talk pages. Both JWPL and WRT have been discussed in chapter 3.6.2.

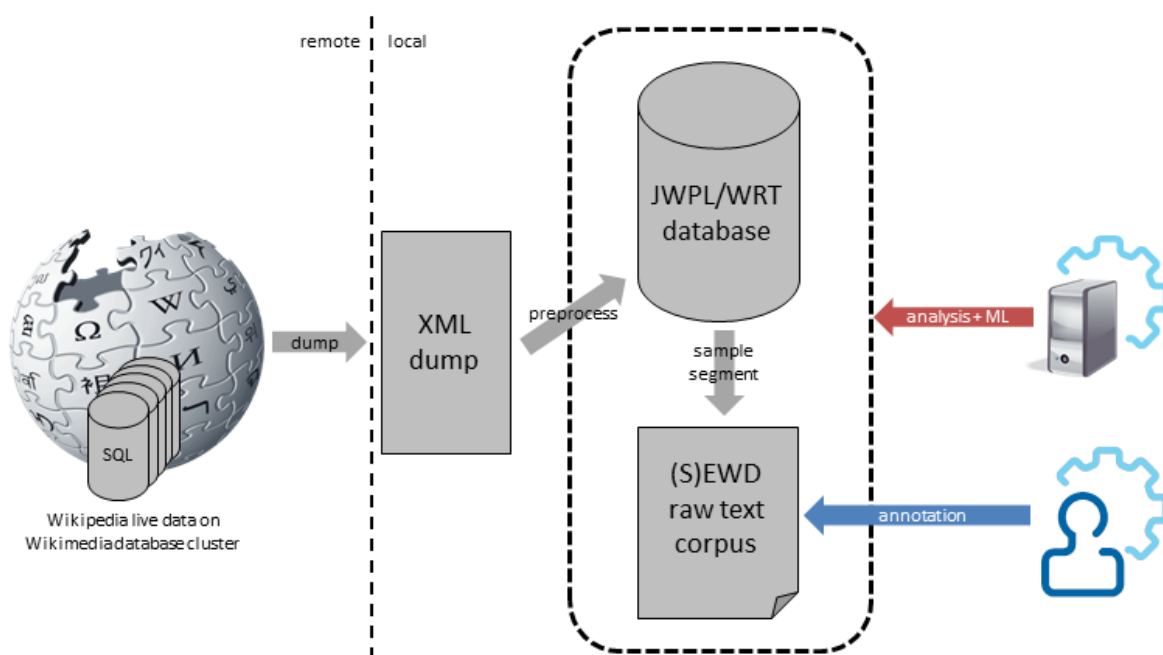


Figure 6.1: Creation and utilization of the Wikipedia Talk page corpora. The corpus creation process is marked in gray, the human annotation task in blue and the computational analysis and machine learning task in red. The XML dumps of the remote Wikipedia databases are provided as downloads by Wikimedia under <http://dumps.wikimedia.org>

For the SEWD corpus, we use a snapshot of the *Simple English Wikipedia*¹¹⁵ from 6th April 2011. For the EWD corpus, we use a snapshot of the *English Wikipedia*¹¹⁶ from 5th April 2011.

6.4.1 Dialog Segmentation

As shown in chapter 3.5, Talk pages are regular wiki pages without any explicit markup of the discourse structure. This lack of structure causes not only considerable confusion and disorientation among the discussing users (Schneider et al., 2011), it also makes automatic processing of these pages challenging.

In order to properly analyze the user discussions, we have to segment the discussion pages and extract the basic discourse structure. We therefore have to (i) identify all discussion threads on the page, (ii) segment each thread into individual turns, and (iii) retrieve meta information for each turn, such as the contributing user and the time stamp of the contribution.

¹¹⁵<http://dumps.wikimedia.org/simplewiki>

¹¹⁶<http://dumps.wikimedia.org/enwiki>

I'll go through and make straightforward changes (please revert if I inadvertently change the meaning) and jot queries below. [Cas Liber \(talk · contribs\)](#) 13:56, 29 September 2013 (UTC)

- Avoid mentioning Cann River twice in the first two sentences.
- *It received its current name in 1958, though the same name had been in use by the Snowy Mountains Highway up until 1955.* - err the same name is Monaro Highway here?
- *In the future the northern end of the Monaro Highway as it currently stands will link onto the Majura Parkway, which is currently under construction* - generally avoid terms suc has "currently" or "recently" - best would be to get an estimated finish/completion date and slot in.
- Has this got a writeup in travel books or mags? If it is a scenic drive it might have....
- Any other information about - dangerous/black spots, traffic accidents, environmental impacts, condition of road, traffic heaviness (in Canberra) - would be good if it could be sourced.....

Well written overall. [Cas Liber \(talk · contribs\)](#) 14:48, 29 September 2013 (UTC)

(a) Original contribution by user Cas Liber

I'll go through and make straightforward changes (please revert if I inadvertently change the meaning) and jot queries below. [Cas Liber \(talk · contribs\)](#) 13:56, 29 September 2013 (UTC)

- Avoid mentioning Cann River twice in the first two sentences.

[Fixed -- Nbound \(talk\) 14:49, 29 September 2013 \(UTC\)](#)
- *It received its current name in 1958, though the same name had been in use by the Snowy Mountains Highway up until 1955.* - err the same name is Monaro Highway here?

[Reworded -- Nbound \(talk\) 14:49, 29 September 2013 \(UTC\)](#)
- *In the future the northern end of the Monaro Highway as it currently stands will link onto the Majura Parkway, which is currently under construction* - generally avoid terms suc has "currently" or "recently" - best would be to get an estimated finish/completion date and slot in.

[Reworded -- Nbound \(talk\) 15:15, 29 September 2013 \(UTC\)](#)
- Has this got a writeup in travel books or mags? If it is a scenic drive it might have....

[Its not known for being scenic, what kind of information were you after? I can certainly attempt to accommodate -- Nbound \(talk\) 15:15, 29 September 2013 \(UTC\)](#)
- Any other information about - dangerous/black spots, traffic accidents, environmental impacts, condition of road, traffic heaviness (in Canberra) - would be good if it could be sourced.....

[I have exams in a week so its unlikely I can get all that in within that time, is there anything you'd prefer more than the rest? -- Nbound \(talk\) 15:15, 29 September 2013 \(UTC\)](#)

Well written overall. [Cas Liber \(talk · contribs\)](#) 14:48, 29 September 2013 (UTC)

[@Casliber: - Thanks for your review, looking forward to hearing your replies. -- Nbound \(talk\) 15:15, 29 September 2013 \(UTC\)](#)

(b) Same contribution with in-text replies

Figure 6.2: In-text replies on a Talk page of the article *Monaro Highway* (English Wikipedia, revision IDs 575007442 and 575010159, emphasis added)

[Laniado et al. \(2011\)](#) and [Danescu-Niculescu-Mizil et al. \(2012\)](#) tackle the dialog segmentation problem by using text indentation and inserted user signatures as clues. However, according to a study by [Viégas et al. \(2007\)](#), only 67% of all contributions on Wikipedia Talk pages are signed, which makes signatures an unreliable predictor for turn boundaries and thus insufficient for a reliable reconstruction of the thread structure. And even from the available signatures, it is not always possible to retrieve the necessary meta information (see figure 3.6 in chapter 3.5).

Another factor that limits the reliability of a rule-based discourse parsing approach is the non-standard usage of the Talk pages. In contrast to conventional threaded discussions, such as web forums or newsgroups, Wikipedia Talk pages might exhibit cases of *in-text replies*, an approach to insert a new message in an existing contribution of a different user in order to reply to a specific part of said contribution. While editing the contributions

of other users is frowned upon according to the Talk page policies, it is not prohibited by technical means and is allowed under special circumstances¹¹⁷. In-text replies are often used as the preferred way to respond to very long turns, especially when they contain multiple questions or action items. The latter can frequently be observed in threads discussing a list of necessary cleanup tasks which have to be addressed before an article can be promoted to *good* or *featured* status. Figure 6.2 shows an example of in-text replies inserted into an existing post that reviews the quality status of the associated article. In the remainder of this section, we describe our approach to reliably segment user discussions.

6.4.1.1 Topic Segmentation

While there is no explicit markup for the inner discourse structure, Talk pages make light use of general purpose MediaWiki markup to provide a rough separation into discussion topics.

Therefore, we can employ a MediaWiki markup parser to identify the outer boundaries of the discussion threads. We use the built-in parser of the JWPL software, which allows retrieving section elements and headlines corresponding to the discussion topics and topic titles on Talk pages. In our experiments, the markup-based boundary identification resulted in a perfect segmentation of the discussion topics. As the only exception, we found that discussion pages of newly created or low-profile articles with little discussion activity do not make use of explicit sectioning. However, as the discussion activity increases, a coarse grained discourse structure emerges automatically.

6.4.1.2 Turn Segmentation

As a second step, we have to identify the turn boundaries within each discussion thread. Despite existing conventions that define how the dialog is supposed to be formatted by the users¹¹⁸, we refrain from making too many assumptions about the format, since the guidelines and best practices are often not being precisely followed by all users. Our only fixed assumption about the discussion format is that every turn starts with a new line. We therefore consider every end-of-line (EOL) character in the discussion text to mark the boundary to a new *paragraph* and every paragraph to be a *turn candidate*. Based on this assumption, we can reliably preprocess the discussion text in order to provide a semi-structured format for our dialog segmentation algorithm.

The main idea of our segmentation algorithm is that the revision history of the discussion page contains all necessary information that is needed to identify author and creation point of each paragraph, which makes it possible to aggregate associated turn candidates to actual turns.

¹¹⁷<http://en.wikipedia.org/wiki/WP:TPO>

¹¹⁸<http://en.wikipedia.org/wiki/WP:TP>

```

Data: unsegmented text of a single discussion  $d$  in talk page  $tp$ 
Result: list of paragraphs with metadata
1  $parlist \leftarrow$  split  $d$  at EOL ;                               /* identify paragraphs */
   /* find creation point of each paragraph                          */
2 foreach  $paragraph\ p$  in  $parlist$  do
   | /* check all talk page revisions starting with the oldest      */
3   | for  $rev \leftarrow tp.oldest$  to  $tp.newest$  do
4   | | if  $rev$  contains  $p$  then /* String matching                */
   | | | /* we found the revision of origin for  $p$                   */
   | | |  $p.author = rev.author$ ; /* collect metadata */
   | | |  $p.timestamp = rev.timestamp$ ;
   | | |  $p.revisionid = rev.revisionid$ ;
   | | | break; /* goto next paragraph */
   | | end
9   | end
10  end
11 end
12 return  $parlist$ 

```

Figure 6.3: Identification of paragraph creation points with forward checking.

We now introduce a naive algorithm which implements the basic idea of this approach, but makes several simplifying assumptions. After that, we identify the problems of the naive algorithms and generalize the simplifications.

Naive Algorithm with Forward Checking. For the naive algorithm, we start by sequentially searching the revision history of the discussion page once for each paragraph to find the revision in which the paragraph first appeared. We define this revision to be the *revision of origin* of that paragraph. The search always moves forward in time, starting with the first (oldest) revision of the discussion page. The paragraph identification is achieved with a simple string matching technique, i.e. the algorithm checks if the current Talk page revision contains the paragraph as a substring. Having identified the revision of origin for each paragraph, we can retrieve the corresponding author, creation point and revision id from the revision metadata. Assuming that each revision produces exactly one turn, i.e. all paragraphs created in a single revision belong to the same turn, we can aggregate all paragraphs with identical revision IDs to single turns and thus retrieve the overall turn structure of the discussion. Figure 6.3 shows the pseudocode for the naive version of the paragraph creation point identification algorithm.

While this naive version illustrates the main idea behind our revision based segmentation approach, it makes several simplifying assumptions that first have to be generalized in order to make the approach applicable to real world problems.

Data: unsegmented text of a single discussion d in talk page tp
Result: list of paragraphs with metadata

```

/* Initialize paragraphs */
1 parlist ← split  $d$  at EOL ; /* identify paragraphs */
2 foreach paragraph  $p$  in parlist do
3   p.startIndex = start position in Talk page;
4   p.endIndex = end position in Talk page;
5   p.contributor = last contributor to Talk page;
6   p.timestamp = timestamp of latest revision of Talk page;
7 end
/* find creation point of each paragraph */
8 foreach paragraph  $p$  in parlist do
9   /* check all talk page revisions starting with the newest */
10  for rev ←  $tp.newest$  to  $tp.oldest$  do
11    foreach DiffAction  $da$  in rev do
12      if  $da$  occurred within limits of  $p$  then
13        /*  $p$  was changed for the first time. update metadata of  $p$  and
14         move to next paragraph */
15         $p.author$  =  $rev.author$ ;
16         $p.timestamp$  =  $rev.timestamp$ ;
17         $p.revisionid$  =  $rev.revisionid$ ;
18        break; goto next paragraph;
19      else
20        /*  $p$  was not changed. recalculate position of  $p$  according to the
21         changes made on the Talk page */
22        update indexes of  $p$  according to  $da$ ;
23      end
24    end
25  end
26 end
27 return parlist

```

Figure 6.4: Identification of paragraph creation points with backward checking.

Backward Checking. One of the simplifications of the naive algorithm is the utilization of string matching for finding the revision of origin of a given paragraph. While this is a viable approach for longer paragraphs, which are most likely unique in the whole revision history, the probability of identifying an incorrect revision of origin increases with decreasing length of the paragraph text. In order to solve this issue, we reverse the processing order of the revision history and start the process with the newest revision of the Talk page. We then backtrack the history revision by revision and check whether the monitored paragraph has been altered. This can be achieved without relying on string matching with the help of the so-called *DiffAction* information provided by the WRT-API (see appendix A.1). These DiffActions identify all changes that have been made in a single revision of a wiki page and provide the span of the change with start and end index in the page. This way, we are able to determine whether any of these changes occur within a particular paragraph by comparing the begin and end indexes of DiffAction and paragraph. The process is continued until we find the first revision in which the monitored paragraph is either altered or disappears completely, which indicates that we found the creation point of the paragraph. Figure 6.4 shows the pseudocode of the creation point identification algorithm with backward checking.

Vandalism. Backward checking introduces another problem that was not relevant in the naive setup. Similar to Wikipedia articles, Talk pages can be subject to vandalism and malicious edits. While there are many possible categories of malicious edits to wiki pages, an extreme example best illustrates the impact of vandalism on our algorithm. *Page blanking* is defined as the action of removing all content from a page or replacing all content on a page with a new message. These acts of vandalism are usually repaired very quickly by someone reverting the malicious change and restoring the previous version of the Talk page. However, such page blanking acts will cause the segmentation algorithm to falsely select the malicious revision as the revision of origin for any paragraph created before the vandalism act. Therefore, we have to account for cases of vandalism by identifying malicious edits. Since we found that vandalism is reverted much faster on discussion pages than in articles, we do not employ any additional vandalism detection heuristics. We rather check whether a change to the currently monitored paragraph is reverted shortly afterwards within a so-called *lookahead window*, i.e. within the next n revisions after the change. In other words, we identify cases of *paragraph blanking* with a n -revision lookahead. If we found such a case, we assume an act of vandalism and disregard the malicious change. We then proceed with the search for the revision of origin for the monitored paragraph. While smaller lookahead windows potentially cause incorrect decisions of the algorithm, larger windows will increase the processing time. On the SEWD corpus, the algorithm was able to detect all cases of paragraph blankings with a lookahead windows of $n = 10$, while we had to increase the value to $n = 20$ on the EWD corpus.

Edit-Turn Equivalence. The final simplification of the naive approach that we have to address is the edit-turn equivalence. We assumed that one revision equals one turn. However, this is not always the case. While in many cases we actually do have a 1:1 correspondence between turns and revisions, we also find cases with 1:N, M:1 or even M:N ratios. This means, that a single turn was written in many revisions, multiple turns in one revision or multiple turns in multiple revisions. We have to consider these cases in the paragraph aggregation part of the algorithm.

In order to account for turns that have been written in multiple revisions, we regard all consecutive revisions by the same user within a window of 10 minutes as belonging to the same turn. The value of this time window was derived experimentally and turned out to be the optimal value for our experiments. We evaluated the threshold in a manual review of all incorrectly segmented turns both in the SEWD and EWD corpus and found that the same threshold worked equally well in both cases.

Multi turn edits, i.e. revisions in which a user contributes to multiple discussion threads on a single page, are not a problem for the revised algorithm, since the paragraph-based revision backtracking approach captures these cases equally well.

Indentation. Even though we do not regard indentation as a reliable indicator for the conversational structure, we nevertheless record this information for every turn in order to gain further insights into the relationships between the contributions. According to the Talk page conventions¹¹⁹, indentation should be used to indicate which part of the conversation a contribution replies to. We store this information as additional metadata for the turn. In cases where a single turn contains indentation as a means of formatting the contribution rather than indicating a reply, the indentation level of the least indented paragraph is used for the whole turn.

In-Text Replies. In the case of in-text replies (see figure 6.2), users insert their messages within an existing contribution. We can consider this to be an act of splitting the original contribution into smaller parts or, in other words, *partial turns*, to which the inserted messages reply. As long as the original contribution is split along paragraph boundaries, our backward checking algorithm will still find the correct creation points for each partial turn. However, if the split is performed somewhere in the middle of a paragraph, we have to account for this fact when monitoring the indexes of the paragraphs. After each index recalculation (line 17), we have to check if the newly calculated span still starts and ends at paragraph boundaries. If this is no longer the case, we have to expand the span to the paragraph boundaries it is enclosed in. Then, the algorithm will be able to perform as expected.

¹¹⁹<http://en.wikipedia.org/wiki/WP:INDENT>

Discussion Archives. As discussed in chapter 3.5, old discussion threads can be archived in order to prevent Talk pages from getting too long. Depending on the configuration of the particular Talk page, the old content is either copied to a new or already existing archive page before being removed from the main page (cut-and-paste procedure) or the whole Talk page is renamed and thus transformed into an archive and a new main Talk page is created (move procedure). In the latter case, the segmentation algorithm works as expected, since the revision history remains unchanged on the same page as the content, i.e. the renamed archive page. In the case of the cut-and-paste procedure, the revision history related to the copied content remains on the main Talk page. Consequently, the algorithm has to use the revision history of the main page starting at the point in time when the content was archived. This point in time can be retrieved from the revision history of the archive page. Since there is no explicit flag indicating the archiving strategy, we initially assume that the move procedure has been employed. If the algorithm fails upon segmenting the page, we automatically switch to the cut-and-paste mode.

6.4.1.3 Evaluation of the Segmentation Approach

In the following, we first evaluate the performance of the segmentation approach in terms of its time complexity and then proceed with an empirical evaluation of its accuracy on the SEWD and the EWD corpus.

Performance Estimation. The main computational effort of the segmentation algorithm lies in the identification of the paragraph creation points.

The naive forward checking algorithm requires, on average, $p * 0.5r$ string matches for p paragraphs and r Talk page revisions. If the Talk page contains the history of archived content, these revisions have to be unsuccessfully searched for every paragraph, since the algorithm always starts the search with the oldest revision.

The backward checking algorithm requires $p(x * r)$ check and update operations.¹²⁰ If the Talk page does not contain any archived revisions, x is 0.5 like in the case of forward checking, because, on average, we have to search half of the history. If any archived revisions are contained in the history of the page, x will be smaller than 0.5, since the backward checking algorithm will never enter the part of the revision history with archived revisions. However, compared to the naive approach, the check and update operations are computationally more complex than string matching. Therefore, the backward checking algorithm will only be faster in cases with many archived discussion threads using the cut-and-paste procedure and thus resulting in smaller values of x .

¹²⁰The check and update operations have to be carried out for each DiffAction in a revision. Since this number is, on average, small, we handle this as a fixed operation. The initialization of the paragraphs in the beginning does not significantly affect the overall runtime, since the properties of every paragraph are assigned with fixed values.

Paragraph aggregation into turns can be achieved in a single iteration over all processed paragraphs if they are first sorted according to creation time. The handling of in-text replies increases the complexity of the check and update operations of the backward checking algorithms, since in-paragraph splits have to be handled as described above. However, in practice, the operations do not affect the overall runtime significantly, since the cases of in-text replies are infrequent compared to standard replies.

Empirical Evaluation of Segmentation Accuracy We evaluated the accuracy of the final segmentation algorithm with backward checking both for the Simple English Wikipedia and the English Wikipedia. We manually analyzed all Talk pages in the SEWD corpus and the EWD corpus for segmentation errors as part of the annotation process described in section 6.5.

In the case of the Simple English Wikipedia, we evaluated on a per-turn basis, which means that each turn was judged individually for the acceptability of its boundaries and the correctness of the associated metadata. For the gold standard of the corpus (see section 6.5.2), any turns with segmentation errors were excluded. Overall, 94% of the 1450 turns were correctly segmented.

An analysis of the incorrectly segmented turns showed two major sources of error. First, the assumption that turns boundaries always coincide with paragraph boundaries (i.e. turns are single or multiple paragraphs) did not always hold true. There are rare cases in which this convention is disregarded by the users and new turns are not started in a new line. Second, the algorithm expects any cases of vandalism to be reverted within a certain time window after it occurred. There are both cases in which the revert is done outside of the lookahead window of our algorithm or in which no revert is performed at all. These cases will not be detected by our approach. Minor sources of error were bot interventions, i.e. automatic edits performed by maintenance scripts, which could not always be handled correctly. Finally, there are cases in which the algorithm fails to keep track of the paragraph spans after recalculating the begin and end indexes according to the performed DiffActions. This is probably caused by inconsistencies in the WRT-API and not an error in the segmentation algorithm.

For the larger EWD corpus, we evaluated the segmentation accuracy on a per-thread basis, which means that each thread with any segmentation error is immediately marked as erroneous and not further evaluated. This was done in order to exclude whole threads from the gold standard rather than removing individual turns. 8% of all threads have been marked erroneous by three annotators who were asked to evaluate the segmentation accuracy (see also section 6.5.2), while 31% of all threads have been marked with an error tag by at least one annotator. Upon later examination, we found that one of the annotators made

Table 6.3: Descriptive statistics for the Simple English Wikipedia (Apr 6th 2011) and the English Wikipedia (Apr 5th 2011), from which the SEWD and the EWD corpora have been extracted. Turn counts have not been determined for the English Wikipedia due to the extensive runtime of the dialog segmentation process.

| | Simple English | English |
|---------------------------|----------------|-----------|
| Articles | 69 900 | 3 477 738 |
| Non-empty Talk pages | 5 783 | 3 353 180 |
| Discussion topics | 7 560 | 2 901 532 |
| Turns | 14 335 | N/A |
| Talk pages with > 3 turns | 683 | N/A |

a systematic error in judging the segmentation accuracy. Therefore, we consider only the ratings of the other two annotators and obtain an overall thread-error-rate of 10%¹²¹.

Besides the fact that longer threads and older discussion pages with a more extensive revision history were more frequently subject to parse errors than short discussions and newly created Talk pages, the interference of automatic maintenance scripts (bots) caused the algorithm to falsely identify turn boundaries on some occasions. Beyond that, no systematic error sources could be identified among the segmentation errors.

6.4.2 Data Sampling

In order to sample a well balanced set of documents for each corpus, we defined selection criteria that reflect both the peculiarities of the respective Wikipedia from which the documents were sampled and the task at hand, i.e. the analysis of coordination efforts for article improvement. Due to the different characteristics of the Simple English Wikipedia and the English Wikipedia, the selection criteria differ between the SEWD corpus and the EWD corpus.

6.4.2.1 Simple English Wikipedia

Due to the limited size of the Simple English Wikipedia (see table 6.3), its smaller community and consequently its lower discussion activity, we chose the discussion length as the only selection criterion for Talk pages to be included in the corpus. From a JWPL database based on a Wikipedia data dump from Apr 6th 2011, we first extract all Talk pages and segment the dialog as described in section 6.4.1. From the set of segmented Talk pages, we discard all instances with less than four turns, which results in a remaining set of 683 Talk pages. We then analyze the distribution of turn counts per discussion page in the remaining set of pages and manually define three classes: (i) discussion pages with 4–10 turns, (ii) pages with 11–20 turns, and (iii) pages with more than 20 turns. We decided to explicitly

¹²¹Thread-error-rate is defined as the percentage of threads with at least one incorrectly segmented turn.

| | |
|-------------------------------|---|
| Distinguished articles | Featured Articles Good Articles |
| Flawed articles | Incomplete articles or articles with lack of detail (CRITCOMPL) – templates from category <i>Cleanup/Expand and add</i> Articles with lack of accuracy, correctness or neutrality (CRITACC) – templates from category <i>Cleanup/Contradiction and confusion</i> – templates from category <i>Cleanup/Neutrality and factual accuracy</i> – template <i>bad summary</i> Articles with deficiencies in language or style (CRITLANG) – templates from category <i>Cleanup/Style of writing</i> – templates from category <i>Cleanup/Translation</i> Articles with deficiencies in structure or layout (CRITSTRUCT) – templates from category <i>Cleanup/Structure, formatting and sections</i> – templates from category <i>Cleanup/Move</i> – templates from category <i>Cleanup/Merge</i> – templates from category <i>Cleanup/Split</i> Articles with unsuitable content (CRITSUIT) – templates from category <i>Cleanup/Potentially unwanted content</i> – templates from category <i>Cleanup/Importance and notability</i> – templates from category <i>Cleanup/Context and detail</i> Articles with insufficient sources or references (CRITAUTH) – templates from category <i>Cleanup/Verifiability and sources</i> |
| Neutral articles | Articles with none of the above characteristics |

Figure 6.5: Selection criteria for Talk pages in the EWD corpus based on the quality status of the associated articles. The template categories are explained in chapter 5 while the templates used in this setup are listed in appendix B.2. The labels in parentheses refer to the corresponding criticism class as defined in the annotation scheme used to annotate the EWD corpus (see section 6.5.1.2).

define these three classes, since random sampling from a small set of documents might exclude rare document types, i.e. in our case longer discussions. We then randomly extracted 50 discussion pages from class (i), 40 pages from class (ii) and 10 pages from class (iii). This way, we obtain 100 Talk pages with a total of 1,450 turns. After removing all segmentation errors, 1,367 turns remain for annotation.

6.4.2.2 English Wikipedia

Since the English Wikipedia, as the biggest of all language versions, offers a much larger amount of data (see table 6.3), we are able to define more complex selection criteria for the documents in the EWD corpus. Rather than only taking the discussion length into account, we also include the quality status of the articles associated with the discussion pages into the set of criteria.

From a JWPL database based on a Wikipedia data dump from Apr 4th 2011, we first extract all Talk pages with at least one discussion and with a total text length between 1,000 and 40,000 characters¹²². We further categorize the retrieved pages according to the quality status of the associated article, which can either be *distinguished*, *flawed*, or *neutral*.

Distinguished articles: *Are marked as featured or good articles (see section 4.3)*

Flawed articles: *Contain cleanup templates that indicate particular quality flaws which correspond to the criticism categories that we introduce in the annotation scheme (see section 6.5.1.2). The concept of quality flaws in Wikipedia has been discussed in chapter 5.*

Neutral articles: *Are neither distinguished nor flawed¹²³.*

For each of these categories, we sample a random set of 72 Talk pages with a balanced distribution of different discussion sizes¹²⁴ resulting in a total of 216 Talk pages. After segmenting the pages with the algorithm described in section 6.4.1, we manually remove all pages with segmentation errors and obtain a corpus of 200 pages with 8,531 turns in 2,689 topics for further annotation. Figure 6.5 gives a more detailed overview of the selection criteria for the EWD corpus.

6.5 Annotating Wikipedia Article Discussions

In this section, we first introduce the two annotation schemes designed for the SEWD and the EWD corpus as well as a mapping between the two. We then describe the annotation process and analyze the annotations in both corpora.

6.5.1 Annotation Schemes for Article Discussions

In section 6.2, we have defined dialog acts as specialized speech acts that identify the function of an utterance in the context of a particular dialog. Rather than performing a fine-grained analysis of the discourse structure in Wikipedia Talk page conversations, we are more interested in the types of quality assessment issues and the coordination efforts for article improvement that are reflected in each contribution. Dialog acts are a suitable tool for this kind of analysis. We chose to perform the dialog analysis on the turn level rather than on the utterance level to avoid the added complexity of an additional utterance segmentation step which is an additional source of noise.

¹²²Since the turn extraction for all discussion pages would demand a substantial amount of time, we preselect candidate pages according to their overall text length (excluding markup). Turn segmentation is performed at a later stage.

¹²³While *neutral* articles do not contain any templates defined in the *flawed* category, they might still exhibit unmarked flaws or contain other types of cleanup templates which are not considered in this setup.

¹²⁴We split the range between 1,000 characters and 40,000 characters per article into six equidistant bins and categorized the articles accordingly. We then sampled the same number of articles from each bin.

Since a single turn may consist of several utterances, it is consequently bound to comprise multiple dialog acts. Therefore, we designed the annotation study as a multi-label classification task, i.e. the annotators can assign one or more labels to each annotation unit while each label is chosen independently. This furthermore reflects the fact that even a single utterance might perform several acts at the same time. For example, the author of the turn

“This part needs to be extended. I will have a look into it later.”

identifies a lack of detail or missing information in the article and, at the same time, self-commits to improving the article later.

The annotation schemes described in the following subsections have been developed in succession, whereas the EWD schemes constitutes a refinement of the SEWD scheme based on the lessons learned in our initial experiments.

6.5.1.1 Simple English Wikipedia

In order to define a scheme for annotating the SEWD corpus, we manually analyzed a random set of thirty Talk pages from the Simple English Wikipedia to identify the types of article deficiencies that are discussed and the way article improvement is coordinated. We first identified four high level categories that need to be considered in an analysis of the information quality management process.

Article Criticism: *Identifies the types of deficiencies in the article. The criticism can refer to the article as a whole or to individual parts of the article.*

Explicit Performative: *Comprises announcements, reports or suggestions of editing activities.*

Information Content: *Captures the purpose of the communication. A contribution can be used to communicate new information to others, to request information, or to suggest changes to established facts.*

Interpersonal: *Refers to the attitude that is expressed towards other participants in the discussion and/or their comments.*

Within each of these four categories, we identified all related speech acts that occurred in the Talk page sample more than once. Moreover, we analyzed the instructions regarding article quality discussions in the Wikipedia Manual of Style in order to identify additional labels for each category. The scheme was iteratively revised on another set of 20 random Talk pages to assure the generalizable nature of the labels identified in the first iteration. The resulting final tagset consists of 17 labels which are listed in table 6.4 along with their respective definitions and an example turn from the SEWD corpus.

| Label | Description | Example |
|------------------------------|---|--|
| Article Criticism | | |
| CM | Content incomplete or lacking detail | <i>It should be added (1) that voters may skip preferences, but (2) that skipping preferences has no impact on the result of the elections.</i> |
| CW | Lack of accuracy or correctness | <i>Kris Kringle is NOT a Germanic god, but an English mispronunciation of Christkind, a German word that means “the baby Jesus”.</i> |
| CU | Unsuitable or unnecessary content | <i>The references should be removed. The reason: The references are too complicated for the typical reader of simple Wikipedia.</i> |
| CS | Structural problems | <i>Also use sectioning, and interlinking</i> |
| CL | Deficiencies in language or style | <i>This section needs to be simplified further; there are a lot of words that are too complex for this wiki.</i> |
| COBJ | Objectivity issues | <i>This article seems to take a clear pro-Christian, anti-commercial view.</i> |
| CO | Other kind of criticism | <i>I have started an article on Google. It needs improvement though.</i> |
| Explicit Performative | | |
| PSR | Explicit suggestion, recommendation or request | <i>This section needs to be simplified further</i> |
| PREF | Explicit reference or pointer | <i>Got it. The URL is http://www.dmbatles.com/history.php?year=1968</i> |
| PFC | Commitment to an action in the future | <i>Okay, I forgot to add that, I’ll do so later tonight.</i> |
| PPC | Report of a performed action | <i>I took and hopefully simplified the “[[en:Prehistoric music Prehistoric music]]” article from EnWP</i> |
| Information Content | | |
| IP | Information providing | <i>“Depression” is the most basic term there is.</i> |
| IS | Information seeking | <i>So what kind of theory would you use for your music composing?</i> |
| IC | Information correcting | <i>In linguistics and generally speaking, when Talking about the lexicon in a language, words are usually categorized as ‘nouns’, ‘verbs’, ‘adjectives’ and so on. The term ‘doing word’ does not exist.</i> |
| Interpersonal | | |
| ATT+ | Positive attitude towards other contributor or acceptance | <i>Thank you.</i> |
| ATTP | Partial acceptance or partial rejection | <i>Okay, I can understand that, but some citations are going to have to be included for [[WP:V]].</i> |
| ATT- | Negative attitude towards other contributor or rejection | <i>Now what? You think you know so much about everything, and you are not even helping?!</i> |

Table 6.4: Annotation scheme for the dialog act classification in Wikipedia discussion pages with examples from the SEWD Corpus. Some examples have been shortened to fit the table.

| Scope | Label | Description |
|--------------------------|------------------------|--|
| Topic | ERROR | Segmentation Errors |
| | REFOBJ-PART | Comment about specific section of the article |
| | REFOBJ-WHOLE | Comment about the whole article |
| | REFOBJ-META | Meta comment not directly referring to article |
| Article Criticism | | |
| Turn | CRITCOMPL | Information is incomplete or lacks detail |
| | CRITACC | Lack of accuracy, correctness or neutrality |
| | CRITLANG | Deficiencies in language and style |
| | CRITSUIT | Content not suitable for an encyclopedia |
| | CRITSTRUCT | Deficiencies in structure, organization or visual appearance |
| | CRITAUTH | Lack of authority |
| | Self Commitment | |
| Turn | ACTF | Commitment to action in the future |
| | ACTP | Report of past action |
| Requests | | |
| Turn | REQEDIT | Request for article edit |
| | REQMAINT | Request for admin or maintenance action |
| Interpersonal | | |
| Turn | ATTPOS | Positive attitude |
| | ATTNEG | Negative attitude |

Table 6.5: Revised annotation scheme for the dialog act classification in the EWD Corpus. Topic level labels are assigned to whole discussion threads while turn level labels are assigned to individual turns.

6.5.1.2 English Wikipedia

Based on the results of the inter-annotator analysis on the SEWD corpus that is discussed in section 6.5.3 and the feedback we received from the annotators, we revised the SEWD scheme before applying it to the EWD corpus.

Most notably, we expanded the annotation scheme with an additional topic scope layer in addition to the turn scope that we already defined for the SEWD experiments. In other words, in addition to the dialog act labels on the turn level, we add an additional annotation layer on the topic level that refers to whole discussion topics. The topic scope labels are intended to provide background information about whole discussion threads. The ERROR label marks segmentation or parse errors in the corpus. If any turn is incorrectly parsed or segmented, the whole thread is marked with an error label and potentially rejected from inclusion in the gold standard. Furthermore, the multi-class REFOBJ label defines the point of reference of the discussion topic. A discussion can either refer to the article as a whole (REFOBJ-WHOLE), a particular section of the article (REFOBJ-PART) or to external content

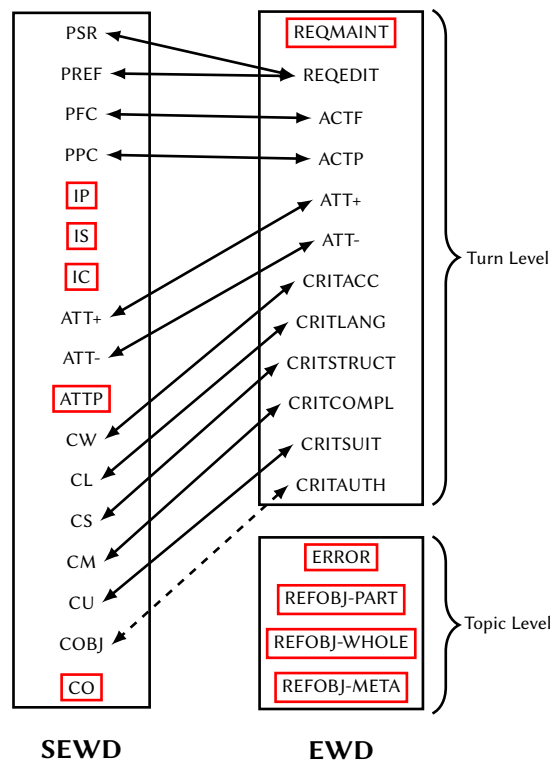


Figure 6.6: Mapping between the SEWD and EWD annotation schemes. Dashed lines indicate partial equivalence of labels. Red borders indicate labels without equivalence in the other scheme.

outside of Wikipedia or an off-topic (REFOBJ-META). For example, a discussion whether the article *Computational Linguistics* should be merged with *Natural Language Processing* refers to the article(s) as a whole. The request to improve the lead section of this article would refer to a particular part. Discussions about current events in natural language processing or general remarks about Wikipedia policies would be regarded as meta discussions. The choice between the three points of reference is mutually exclusive, hence the REFOBJ label was defined as a multi-class label instead of three binary labels. That is, a discussion topic can either be marked with REFOBJ-WHOLE, REFOBJ-PART or REFOBJ-META.

In particular, the very unspecific CO label was removed due to very low inter-annotator agreement. We furthermore merged the PREF label with the PSR label into a single REQEDIT label, since the two have frequently been confused. We additionally introduced the REQMAINT label indicating requests for maintenance activities that can only be carried out by privileged users such as administrators.

All labels from the *Information Content* category have been discarded, because the IP label has shown to be too unspecific while the other labels suffered either from low frequency or from low inter-annotator agreement. We finally removed the ATTP label that originally expressed partial agreement between the users in a discussion. Instead, we defined partial agreement to be represented by assigning both the labels for positive and negative attitude.

To improve the overall inter-annotator agreement and to take the increased complexity of the longer discussions into account, we created a more comprehensive annotator’s manual that was particularly tailored towards providing guidelines for unclear cases. An abridged version of this manual can be found in appendix B.2.

Figure 6.5 shows the revised annotation scheme for the EWD corpus while figure 6.6 demonstrates how both annotation schemes can be mapped to each other and which of the labels do not have an equivalent in the other scheme.

6.5.2 Corpus Annotation Process and Gold Standard Creation

We use the MMAX2 annotation tool (Müller and Strube, 2006) for annotating both the SEWD and the EWD corpus. Therefore, we convert the segmented Talk pages into the MMAX2 XML format. For the SEWD corpus, we define all turn boundaries as markables, i.e. units annotatable according to the predefined annotation scheme, whereas we define two markable levels for the EWD corpus based on turn and topic boundaries.

The screenshot in figure 6.7 shows the MMAX2 annotation tool being used for labeling an in-text reply, i.e. a turn inserted into an existing turn by another user. The inserted turn is highlighted in yellow marking it ready for annotation while the turn in which the in-text reply was inserted becomes a discontinuous turn and is marked in gray. Both parts of the discontinuous turn are still connected and can be annotated as one unit.

Figure 6.8 shows the topic annotation screen. All turns in the topic are highlighted while the topic metadata is displayed in the annotation window.

Gold Standard Creation. The SEWD gold standard was created by a third expert annotator who manually consolidated the annotation of the two trained annotators. This consolidation process was carried out within the MMAX2 system that was already used in the annotation process.

In order to improve the consolidation process for the bigger EWD corpus and make it easier to handle a larger amount of annotations, we designed a dedicated expert support system. We first read the annotations of both annotators into a UIMA pipeline and store them in individual annotation layers in the same CAS¹²⁵. In a second step, we remove all discussion threads that have been marked with an ERROR label by any annotator. While this potentially discards usable data, we achieve the highest possible accuracy and reliability. This way, we obtain an overall number of 4,884 turns not marked with any error. In a third step, we identify all cases of disagreement between the annotators and mark them as separate annotations in the CAS. Cases with perfect agreement between the annotators are directly accepted by the system thus rendering additional review by the expert unnec-

¹²⁵Common Analysis Structure, the object based data structure of the UIMA framework that holds both the data and any standoff annotations.

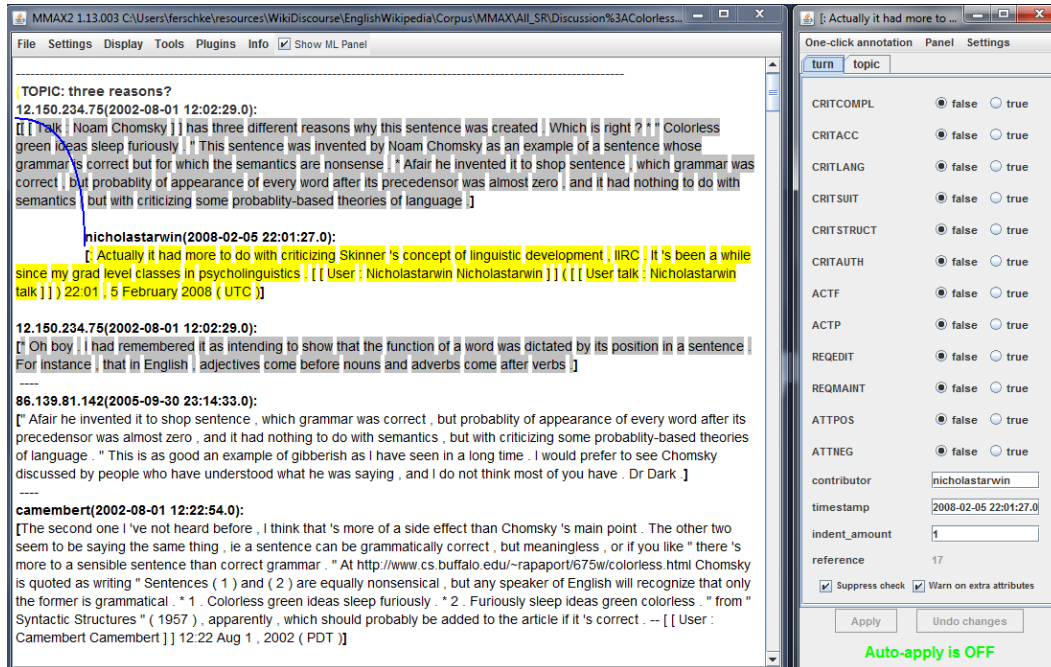


Figure 6.7: Annotation of a Talk page from the EWD corpus on the *turn level* in MMAX2. The currently selected turn is highlighted in yellow. A blue line indicates a reply to a previous turn. In this example, the yellow turn is an inserted reply which was placed within a previous turn (marked in gray) that was thereby split in two parts. The turn-level labels from the annotation scheme are displayed on the right.

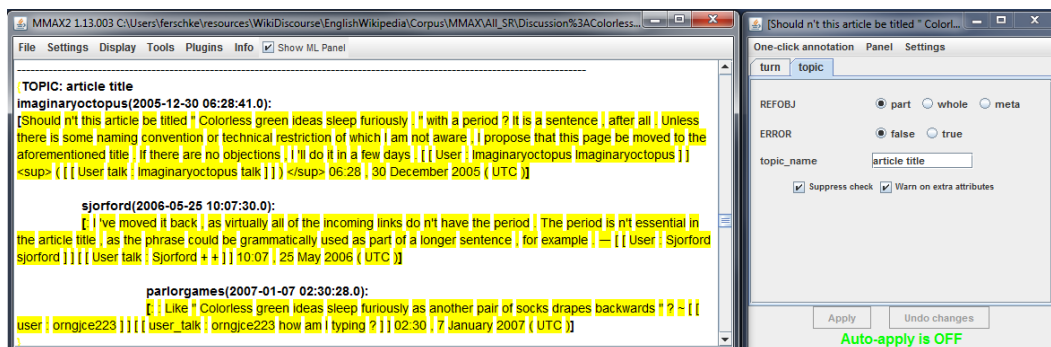


Figure 6.8: Annotation of a Talk page from the EWD corpus on the *topic level* in MMAX2. All turns belonging to the currently selected topic are highlighted in yellow. The labels from the annotation scheme are displayed on the right.

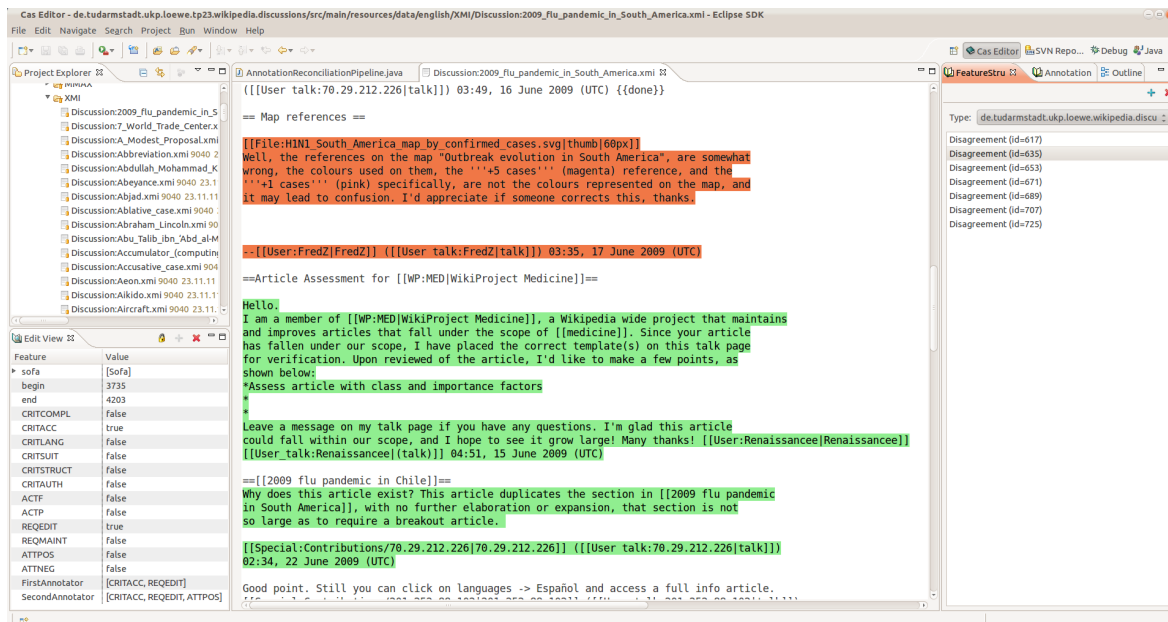


Figure 6.9: UIMA CasEditor as the front end of the expert support system for creating the EWD gold standard.

essary. Using the Apache UIMA CasEditor¹²⁶, the expert annotator can then navigate the disagreement annotations, review the annotator decisions, and enter the final gold standard labels. The resulting gold standard corpus can finally be saved in the UIMA XMI¹²⁷ format, which allows further processing with the UIMA framework.

6.5.3 Inter-Annotator Agreement

To evaluate the reliability of our datasets, we perform a detailed inter-rater agreement study. For measuring the agreement of the individual labels, we report the observed agreement, Kappa statistics (Carletta, 1996), and F_1 -scores. The latter are computed by treating one annotator as the gold standard and the other one as predictions (Hripcsak and Rothschild, 2005). The detailed scores for the SEWD corpus along with descriptive statistics regarding label assignments are shown in table 6.6 while the EWD corpus is summarized in table 6.7. A breakdown of the dialog act label assignments per topic is furthermore presented in table 6.8.

For the SEWD corpus, the average observed agreement across all labels is $P_0^- = 0.94$. The individual Kappa scores largely fall into the range that Landis and Koch (1977) regard as *substantial agreement*, while three labels are above the more strict 0.80 threshold for

¹²⁶<http://uima.apache.org/downloads/releaseDocs/2.3.0-incubating/docs/html/tools/tools.html#ugr.tools.ce> accessed on Feb 20, 2014

¹²⁷<http://uima.apache.org/downloads/releaseDocs/2.3.0-incubating/docs/html/references/references.html#ugr.ref.xmi> accessed on Feb 20, 2014

| Label | Annotator 1 | | Annotator 2 | | Inter-Annotator Agreement | | | | Gold Standard | |
|------------------------------|-------------|---------|-------------|---------|---------------------------|-------|----------|-------|---------------|---------|
| | N | Percent | N | Percent | $N_{A_1 \cup A_2}$ | P_O | κ | F_1 | N | Percent |
| Article Criticism | | | | | | | | | | |
| CM | 183 | 13.4% | 105 | 7.7% | 193 | .93 | .63 | .66 | 116 | 8.5% |
| CW | 106 | 7.8% | 57 | 4.2% | 120 | .95 | .52 | .55 | 70 | 5.1% |
| CU | 69 | 5.0% | 35 | 2.6% | 83 | .95 | .38 | .40 | 42 | 3.1% |
| CS | 164 | 12.0% | 101 | 7.4% | 174 | .94 | .66 | .69 | 136 | 9.9% |
| CL | 195 | 14.3% | 199 | 14.6% | 244 | .93 | .73 | .77 | 219 | 16.0% |
| COBJ | 27 | 2.0% | 23 | 1.7% | 29 | .99 | .84 | .84 | 27 | 2.0% |
| CO | 20 | 1.5% | 59 | 4.3% | 71 | .95 | .18 | .20 | 48 | 3.5% |
| Explicit Performative | | | | | | | | | | |
| PSR | 458 | 33.5% | 351 | 25.7% | 503 | .86 | .66 | .76 | 406 | 29.7% |
| PREF | 43 | 3.1% | 31 | 2.3% | 51 | .98 | .61 | .62 | 45 | 3.3% |
| PFC | 73 | 5.3% | 65 | 4.8% | 86 | .98 | .76 | .77 | 77 | 5.6% |
| PPC | 357 | 26.1% | 340 | 24.9% | 371 | .97 | .92 | .94 | 358 | 26.2% |
| Information Content | | | | | | | | | | |
| IP | 1084 | 79.3% | 1027 | 75.1% | 1135 | .89 | .69 | .93 | 1070 | 78.3% |
| IS | 228 | 16.7% | 208 | 15.2% | 256 | .95 | .80 | .83 | 220 | 16.1% |
| IC | 187 | 13.7% | 109 | 8.0% | 221 | .89 | .46 | .51 | 130 | 9.5% |
| Interpersonal | | | | | | | | | | |
| ATT+ | 71 | 5.2% | 140 | 10.2% | 151 | .94 | .55 | .58 | 144 | 10.5% |
| ATTP | 71 | 5.2% | 30 | 2.2% | 79 | .96 | .42 | .44 | 33 | 2.4% |
| ATT- | 67 | 4.9% | 74 | 5.4% | 100 | .96 | .56 | .58 | 87 | 6.4% |

Table 6.6: Label frequencies and inter-annotator agreement for the SEWD corpus. $N_{A_1 \cup A_2}$ denotes the number of turns that have been labeled with the given label by at least one annotator. P_O denotes the observed agreement.

reliable annotations (Artstein and Poesio, 2008). Furthermore, we obtain an overall pooled Kappa (De Vries et al., 2008) of $\kappa_{pool} = 0.67$, which is defined as

$$\kappa_{pool} = \frac{P_O^- - P_E^-}{1 - P_E^-}$$

with

$$P_O^- = \frac{1}{L} \sum_{l=1}^L P_{O_l} \quad , \quad P_E^- = \frac{1}{L} \sum_{l=1}^L P_{E_l}$$

where L denotes the number of labels, P_{E_l} the expected agreement and P_{O_l} the observed agreement of the l^{th} label. κ_{pool} is regarded to be more accurate than the averaged Kappa.

| Label | Annotator 1 | | Annotator 2 | | Inter-Annotator Agreement | | | | Gold Standard | |
|--------------------------|-------------|---------|-------------|---------|---------------------------|-------|----------|-------|---------------|---------|
| | N | Percent | N | Percent | $N_{A_1 \cup A_2}$ | P_O | κ | F_1 | N | Percent |
| Article Criticism | | | | | | | | | | |
| CRITCOMPL | 323 | 6.6% | 404 | 8.3% | 501 | .94 | .59 | .62 | 373 | 7.6% |
| CRITACC | 671 | 13.8% | 603 | 12.4% | 842 | .92 | .63 | .64 | 605 | 12.4% |
| CRITLANG | 233 | 4.8% | 234 | 4.8% | 330 | .96 | .57 | .58 | 235 | 4.8% |
| CRITSUIT | 457 | 9.4% | 293 | 6.0% | 579 | .92 | .41 | .51 | 321 | 6.6% |
| CRIT- STRUCT | 311 | 6.4% | 329 | 6.8% | 467 | .94 | .51 | .55 | 294 | 6.0% |
| CRITAUTH | 306 | 6.3% | 369 | 7.6% | 498 | .93 | .49 | .62 | 315 | 6.4% |
| Self Commitment | | | | | | | | | | |
| ACTF | 244 | 5.0% | 276 | 5.7% | 352 | .96 | .63 | .63 | 221 | 4.5% |
| ACTP | 681 | 14.0% | 652 | 13.4% | 810 | .94 | .75 | .71 | 551 | 11.3% |
| Requests | | | | | | | | | | |
| REQEDIT | 518 | 10.6% | 1024 | 21.0% | 1178 | .83 | .39 | .60 | 419 | 8.6% |
| REQMAINT | 23 | 0.5% | 79 | 1.6% | 98 | .89 | .07 | .10 | 13 | 0.3% |
| Interpersonal | | | | | | | | | | |
| ATTPOS | 452 | 9.3% | 529 | 10.9% | 646 | .94 | .65 | .66 | 457 | 9.4% |
| ATTNEG | 200 | 2.9% | 143 | 2.9% | 254 | .97 | .50 | .36 | 206 | 4.2% |

Table 6.7: Label frequencies and inter-annotator agreement for the EWD corpus. $N_{A_1 \cup A_2}$ denotes the number of turns that have been labeled with the given label by at least one annotator. P_O denotes the observed agreement.

For assessing the overall inter-rater reliability of the label set assignments *per turn*, we chose Krippendorff’s Alpha (Krippendorff, 1980) using MASI, a measure of agreement on set-valued items, as the distance function (Passonneau, 2006). MASI accounts for partial agreement if the label sets of both annotators overlap in at least one label. We achieved an Alpha score of $\alpha = 0.75$ on the SEWD corpus. According to Krippendorff, datasets with this score are considered reliable and allow tentative conclusions to be drawn.

For the EWD corpus, the average observed agreement across all labels is $P_O^- = 0.94$, similar to the results we achieved in the experiments on the SEWD corpus. However, the individual Kappa scores are generally lower and largely fall into the range that Landis and Koch (1977) regard as *moderate agreement* except for three labels on which the annotators achieved substantial agreement. Overall, we obtain a pooled Kappa of $\kappa_{pool} = 0.55$. Krippendorff’s Alpha measuring the turn-level agreement shows an overall value of $\alpha = 0.57$ using again MASI as a distance metric. Since the *requests* category held labels with particularly low agreement, the reasons for which we discuss later in this section, we also calculated

the overall agreement for all labels except REQMAINT and REQEDIT. After excluding the requests category, we obtain an Alpha of $\alpha = 0.60$ and a pooled Kappa of $\kappa_{pool} = 0.59$.

In the SEWD corpus, the CO label showed the lowest agreement of only $\kappa = 0.18$. The label was supposed to cover any criticism that is not covered by a dedicated label. However, the annotators reported that they chose this label when they were unsure whether a particular criticism label would fit a certain turn or not.

Labels in the interpersonal category all show agreement scores below 0.60. It turned out that the annotators had a different understanding of these labels. While one annotator assigned the labels for any kind of positive or negative sentiment, the other one used the labels to express agreement and disagreement between the participants of a discussion.

In the EWD corpus, the two labels in the *request* category show the lowest agreement of 0.39 (REQEDIT) and 0.07 (REQMAINT) respectively. The low agreement on the latter label was mainly due to the very low frequency of the maintenance requests in the dataset. The former label, on the other hand, was frequently not recognized in turns that both contain a type of self commitment and a request or in which the request is less pronounced. For example, the turn

Ok, I've restored the links, even the com is legitimate. Feel free to remove any link that you think is spam, but please explain. Thanks. [[User:Pm master|Pm master]]
23:17, 19 July 2007 (UTC)

should have both been marked with an ACTP label, due to the restored links, and with a REQEDIT label, due to the suggestion to review the list and remove any links that do not fit. Cases like this, with more subtle requests, have often not been assigned correctly with this label. Furthermore, annotator 2 had the tendency to falsely interpret ordinary questions as edit requests thus causing twice as many assignments of REQEDIT labels as annotator 1 and many false positives.

The relatively high inter-annotator agreement on the assignment of the ACTP label was due to the clear lexical cues associated with this category. The annotators reported that they were mainly looking for terms like *I edited*, *I removed*, or *I revised* to decide whether or not to assign this label. This also explains why the prediction performance of the machine learning classifier described later in this chapter is the highest for this label. Interestingly, the assignment of the similar ACTF label, which indicates future commitment, could less reliably be decided based on lexical cues.

A common problem for all labels in both corpora were contributions with a high degree of indirectness and implicitness. Indirect contributions have to be interpreted in the light of conversational implicature theory (Grice, 1975), which requires contextual knowledge for decoding the intentions of a speaker. For example, the message

Is population density allowed to be n/a?

has the surface form of a question. However, the context of the discussion revealed that the author tried to draw attention to the missing figure in the article and requested it to be filled or removed. The annotators rarely made use of the context, which was a major source for disagreement in the study.

Another difficulty for the annotators were long discussion turns. While, in the SEWD corpus, the average turn consists of 42 tokens, the largest contribution in the corpus is 658 tokens long. In the EWD corpus, this problem was even more severe, with the average turn length being 108 tokens and the longest contribution to span 3,344 tokens. Turns of this size can cover many topics and subjects and thus comprise many different dialog acts, which increases the probability of disagreement among the annotators. As we have mentioned before, we initially decided to annotate the corpus on the turn level since we were mainly interested in a coarse-grained, turn-based dialog act analysis to identify the types of quality assessment issues and the coordination efforts for article improvement that are reflected in each contribution. Given the low inter-annotator agreement on discussions with long turns, the markables should be reduced to smaller units in order to simplify the individual decisions the annotators have to make and thus improve the overall agreement. This can, after all, be addressed by going from the turn level to the utterance level in future work, because individual utterances are much more limited in their goals and actions than whole turns, which makes consistent annotation easier. This, however, will involve additional efforts for discourse parsing, because the turns have further be segmented into utterances, which introduces additional noise to the already error prone parsing process that we introduced in this chapter.

A comparison of our results with the agreement reported for other datasets shows that the reliability of our annotations lies well within the field of the related work. [Bender et al. \(2011\)](#) carried out an annotation study of social acts in 365 discussions from 47 Wikipedia Talk pages. They report Kappa scores for thirteen labels in two categories ranging from 0.13 to 0.66 per label. The overall agreement for each category was $\kappa = 0.50$ and $\kappa = 0.59$, respectively, which is considerably lower than our $\kappa_{pool} = 0.67$, but comparable to the agreement our annotators achieved on the EWD corpus. [Kim et al. \(2010b\)](#) annotate pairs of posts taken from an online forum. They use a dialog act tagset with twelve labels customized for modeling troubleshooting-oriented forum discussions. For their corpus of 1,334 posts, they report an overall Kappa of 0.59. [Kim et al. \(2010a\)](#) identify unresolved discussions in student online forums by annotating 1,135 posts with five different speech acts. They report Kappa scores per speech act between 0.72 and 0.94. Their better results might be due to a more coarse grained label set.

6.5.4 Corpus Analysis

In the following, we provide an analysis of both the SEWD and the EWD gold standard (see section 6.5.2).

6.5.4.1 SEWD Corpus

The SEWD corpus contains 313 discussions consisting of 1,367 turns by 337 users. The average length of a turn is 42 words. 208 of the 337 contributors are registered Wikipedia users, 129 wrote anonymously. On average, each contributor wrote 168 words in 4 turns. However, there was a cluster of 16 people with ≥ 20 contributions.

Table 6.6 shows the frequencies of all labels in the SEWD corpus. The most frequent labels are *information providing* (IP), *requests* (PSR) and *reports of performed edits* (PPC). The IP-label was assigned to more than 78% of all 1367 turns, because almost every contribution provides a certain amount of information. The label was only omitted if a turn merely consisted of a discussion template but did not contain any text or if it exclusively contained questions.

More than a quarter of the turns are labeled with PSR and PPC, respectively. This indicates that edit requests and reports of performed edits are the main subject of discussion. Generally, it is more common that edits are reported after they have been made than to announce them before they are carried out, as can be seen in the ratio of PPC to PFC labels. The number of turns labeled with PSR is almost the same as the number of contributions labeled with either PPC or PFC. This allows the tentative conclusion that nearly all requests potentially lead to an edit action. As a matter of fact, the most common label adjacency pair¹²⁸ in the corpus is PSR→PPC, which substantiates this assumption.

Article criticism labels have been assigned to 39.4% of all turns. Almost half (241) of the labels from this class are assigned to the first turn of a discussion. This shows that it is common to open a discussion in reference to a particular deficiency of the article. The large number of CL labels compared to other labels from the same category is due to the fact that the Simple English Wikipedia requires authors to write articles in a way that they are understandable for non-native speakers of English. Therefore, the use of adequate language is one of the major concerns of the Simple English Wikipedia community.

6.5.4.2 EWD Corpus

The EWD corpus contains 1,864 discussions consisting of 4,923 turns by 2,438 users. The average length of a turn is 109 tokens. 1,682 of the 2,438 contributors are registered Wikipedia users while 750 wrote anonymously. On average, each contributor produced 220 tokens in 2 turns while the user with the most contributions in the corpus produced 102

¹²⁸A label transition $A \rightarrow B$ is recorded if two adjacent turns are labeled with A and B , respectively.

| Label | REFOBJ | | |
|------------------|--------|-------|------|
| | WHOLE | PART | META |
| CRITCOMPL | 43 | 330 | – |
| CRITACC | 62 | 539 | 4 |
| CRITLANG | 5 | 230 | – |
| CRITSUIT | 42 | 278 | 1 |
| CRITSTRUCT | 77 | 216 | 1 |
| CRITAUTH | 33 | 282 | – |
| ACTF | 29 | 189 | 3 |
| ACTP | 84 | 464 | 3 |
| REQEDIT | 56 | 363 | – |
| REQMAINT | 4 | 9 | – |
| ATTPOS | 83 | 373 | 1 |
| ATTNEG | 38 | 157 | 11 |
| Total topics | 130 | 1,687 | 47 |
| Unlabeled topics | 13 | 340 | 39 |

Table 6.8: Distribution of dialog acts in the EWD corpus broken down according to the scope of the topic they occur in as defined by the REFOBJ label. It identifies if a discussion refers to the whole article (WHOLE), a specific part of the article (PART) or to content outside of Wikipedia (META). Unlabeled topics do not contain any turns labeled with any dialog act label from our annotation scheme.

turns and a total of 17,304 tokens. While, in the SEWD corpus, every turn received at least one dialog act label due to the universal applicability of the labels in the information content category, the EWD corpus contains only 2,729 labeled turns. No dialog act labels were applicable to the remaining 2,194 turns. This was to be expected after redesigning the annotation scheme since not every turn contributed to article quality assessment and improvement activities.

Across all four high-level dialog act categories, criticism labels have been assigned most frequently to turns with overall 1,778 turns being marked with at least one of these labels. Self commitments could be identified in 749 turns while 432 turns request either article edits or maintenance activities. Finally, positive or negative sentiment or attitudes towards other participants in the discussion are expressed in 655 turns. Table 6.7 shows the frequencies of all individual labels in the EWD corpus.

Most of the discussion topics in the EWD corpus refer to a specific aspect of the associated article as identified by the REFOBJ label. While 130 discussions refer to the article as a whole, i.e. the concept represented by the article, 1,687 discussions refer to specific sections in the article. This reflects the expected usage of the Talk pages, since the work coordination that can be observed there mainly focuses on small, fine-grained steps rather than big picture discussions. 47 discussions are about topics not directly related to the article and mainly discuss general Wikipedia policies. Table 6.8 gives a breakdown of dialog acts according to the scope of the topic they occur in.

6.6 Automatic Prediction of Dialog Act Labels

As we have already discussed in section 6.5.1, dialog act classification is a multi label classification task. That is, given a turn $t \in T$ and a set of dialog act labels $C = \{c_1, c_2, \dots, c_n\}$, we want to label each turn t with $L \subset C$, where L is the set of relevant or true labels and $|L| \geq 1$. Each label is considered to be binary, thus we have $|C| = 2$. According to Tsoumakas and Katakis (2007), multi label classification problems can be approached in two different ways, either by adapting single label learning algorithms to directly support multi labeled training data and thus incorporate label interdependencies in the training and classification process (*algorithm adaptation*) or by decomposing the classification problem into multiple single label problems (*problem transformation*). In our experience, decomposing the multi label problem into individual binary classification problems is the superior solution for noisy data and performs better than algorithms that tackle the multi label classification in its full form. We therefore choose the problem transformation approach for our experiments. It allows us to train individual classifiers for each label which can either be employed separately or combined in an ensemble method (Fujino et al., 2008). The sequential nature of the dialog will furthermore be explicitly reflected in the feature set.

In the following, we first describe the setup of our classification system, present the features employed in our experiments and finally evaluate the performance of the classifiers including an error analysis.

6.6.1 Experiment and System Setup

Similar to our approach in the quality flaw prediction experiments described in chapter 5, we developed a UIMA-based (Ferrucci and Lally, 2004) text classification system using the Weka data-mining software (Hall et al., 2009) as a downstream machine learning toolkit. In contrast to the FlawFinder system, which is organized in several independent and self-sustained *processing tasks*, the dialog act classification system consists of a single UIMA pipeline containing all preprocessing and classification components.

Preprocessing. All necessary preprocessing steps have already been carried out during the dialog segmentation process described in section 6.4.1. The discussions of the gold standard corpus are already segmented into discussion topics, turns, sentences and tokens and basic meta information about the contributors and their contributions are provided in the corpus. Following the hybrid corpus approach described in section 6.4, additional information can be accessed via the JWPL database containing the full Talk page revision history as well as the history of the associated article.

Classification Algorithms. For the classification task, we use three machine learning algorithms from the Weka data-mining software that have proven to work particularly well for similar tasks that we described in related work. We use a Naive Bayes classifier, J48, an implementation of the C4.5 decision tree algorithm (Quinlan, 1992) and SMO, an optimization algorithm for training support vector machines (Platt, 1998). We only employed the default configurations for each machine learning algorithm as defined by the Weka software and did not perform hyperparameter optimization since we were more interested in the feature engineering aspects rather than tweaking the configurations of the algorithms. However, we evaluate the performance of each learning algorithm separately on every dialog act label in order to identify the best classifier combination for the final ensemble pipeline.

Handling Class Imbalance. Since the number of positive instances for each label is small compared to the number of negative instances, we create a balanced dataset which contains an equal amount of positive and negative instances. Therefore, we randomly select the appropriate number of negative instances and discard the rest. This way, we avoid the classifier to be biased towards the majority class. A similar effect can also be reached with cost-sensitive learning (Ling and Sheng, 2010). However, we found that undersampling of the majority class achieves similar results while improving the training speed and is also superior to oversampling the minority class.

Feature Selection. We evaluated two different feature selection approaches to prune the feature space and select the most meaningful features for each label, Information Gain (Mitchell, 1997) and the χ^2 metric (Yang and Pedersen, 1997), but we did not see systematic differences between the two. Even though we first attempted to select a suitable feature selection approach separately for each label we finally decided to use χ^2 at all times. We now give an overview of the feature types employed in our experiments.

6.6.2 Features

Since we expect lexical cues to be among the most prominent features for the dialog act classification task, we employ token uni-, bi- and trigrams that occurred in at least three different turns of the corpus. Similar to the flaw prediction experiments described in chapter 5, we replace all links to external pages with a generic EXTERNALLINK label while we mark wiki-internal links with an INTERNALLINK label. We furthermore perform stopword filtering using the stopword list from the snowball stemmer¹²⁹, which we augmented with punctuation marks.

¹²⁹<http://snowball.tartarus.org/algorithms/english/stop.txt>

Dialog has a sequential nature, i.e. the probability of a particular turn being tagged with a specific dialog act label depends on the labels of the previous turn and influences the succeeding turn. This aspect can be accounted for by following a dedicated sequence classification approach. Instead of restricting our system to sequence classifiers, we instead chose to incorporate the sequential information on the feature level. To this end, each turn is not only represented by the ngrams extracted from its own text but also includes the ngrams of the previous and the next turn. This way, the preceding and succeeding turn influences the label assignment of the given turn.

Since user discussions are likely to suffer from spelling mistakes, the quality and predictive power of the lexical features can be improved by incorporating spelling error correction in a preprocessing step before feature extraction. We did not include this in our system but strongly suggest to do so in future work.

Besides lexical features, we included surface information, such as the length of the current, previous and next turn (in tokens), temporal information, such as the time distance of a turn to the previous and the next turn (in seconds), and positional information, such as the position of a turn within the discussion, its indentation level and two binary features indicating whether a turn references or is referenced by another turn. We assume that a turn t_2 references a preceding turn t_1 if the indentation level of t_2 is one level deeper than that of t_1 .

Indentation and *temporal distance to the preceding turn* proved to be the best ranked non-lexical features overall. Additionally, the *turn position within the topic* was a crucial feature for most labels in the criticism class and for the label PSR respective REQEDIT. This is not surprising, because article criticism and suggestions respective requests tend to occur in the beginning of a discussion. The two *reference* features have not proven to be useful, since the relational information was already covered by the *indentation* feature. The subjective quality of the lexical features seems to be correlated with the inter-annotator agreement of the respective labels. Features for labels with low agreement contain many n-grams without any recognizable semantic connection to the label. For labels with good agreement, the feature lists almost exclusively contain meaningful lexical cues.

6.6.3 Evaluation and Error Analysis

We used the SEWD corpus as a test bed for identifying the best system configuration which we then also apply to the larger EWD corpus using the label mapping shown in figure 6.6. On the EWD corpus, we only train classifiers for labels on the turn level, but use the topic level labels as a filter to only include turns about the article content while discarding mere meta discussions (REFOBJ-META).

Table 6.9 gives an overview of the performance of all learning algorithms per label on the SEWD corpus as well as the performance of the final ensemble classification pipeline

| Label | Naive Bayes | J48 | SMO | Best |
|----------------------|-------------|-----|-----|------|
| CM | .68 | .48 | .66 | .68 |
| CW | .70 | .20 | .56 | .70 |
| CU | .66 | .35 | .59 | .66 |
| CS | .67 | .67 | .75 | .75 |
| CL | .70 | .66 | .73 | .73 |
| COBJ | .78 | .51 | .63 | .78 |
| CO | .61 | .06 | .39 | .61 |
| PSR | .72 | .70 | .76 | .76 |
| PREF | .76 | .41 | .64 | .76 |
| PFC | .70 | .62 | .73 | .73 |
| PPC | .74 | .82 | .85 | .85 |
| IP | .83 | .93 | .93 | .93 |
| IS | .79 | .86 | .85 | .86 |
| IC | .67 | .32 | .59 | .67 |
| ATT+ | .61 | .65 | .72 | .72 |
| ATTP | .72 | .25 | .62 | .72 |
| ATT- | .52 | .30 | .52 | .52 |
| Macro average | .70 | .52 | .68 | .73 |
| Micro average | .74 | .75 | .80 | .82 |

Table 6.9: F_1 -Scores for all classifiers trained on the balanced dataset from the SEWD-corpus obtained with 10-fold cross-validation. *Best* refers to our final ensemble classification pipeline.

evaluated on 10-fold cross validation. Naive Bayes performed surprisingly well and showed the best macro averaged scores among the three learners while SMO showed the best micro averaged performance. We furthermore compare our results to two random baselines and to the performance of the human annotators (cf. figure 6.10). While *baseline 1* assigns labels according to their frequency distribution in the unbalanced dataset, *baseline 2* assigns the labels randomly on the balanced dataset. Our final classifier outperformed the baselines on all labels.

The comparison with the human performance shows that our system is able to reach the human performance. In most cases, the annotation agreement is reliable, and so are the results of the automatic classification. For the labels CU and CO, the inter-annotator agreement is low. The comparatively good performance of the classifiers on these labels shows that the instances do have shared characteristics that make automatic classification possible but they might not be salient enough for human raters to pick up on in manual annotation.

On the EWD corpus, we employ the best configuration obtained on the SEWD corpus. Due to the small number of instances available for the REQMAINT label, we exclude it from

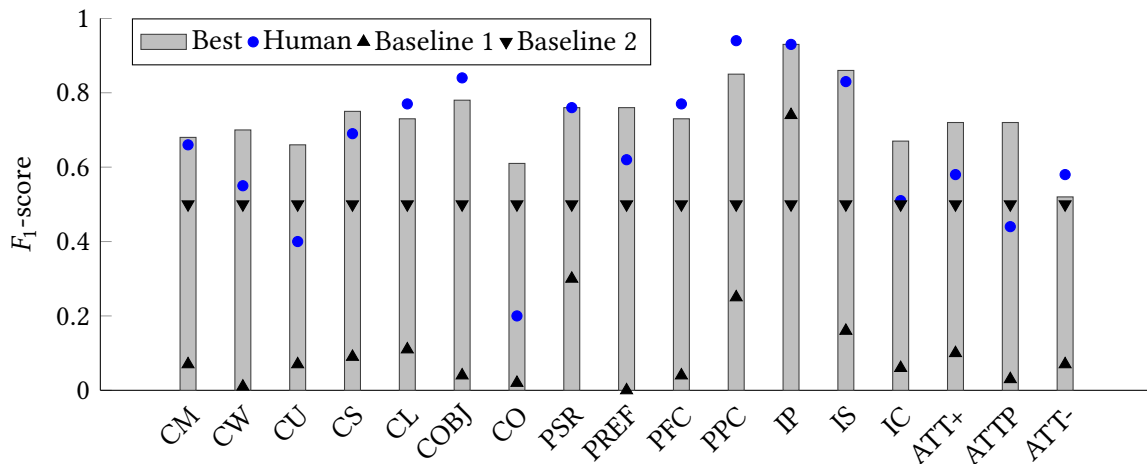


Figure 6.10: F_1 -Scores for the classification pipeline (*Best*), the human performance and baseline performance on the SEWD corpus. Baseline 1 assigns labels according to their frequency distribution in the unbalanced dataset, while baseline 2 assigns labels at random on the balanced dataset.

the experiments, since the amount of data is not sufficient for training a classifier with it. While the classifiers trained on the full dataset only achieve an average performance of $F_1 = 0.56$, training on a balanced dataset improves the performance to an average of $F_1 = 0.78$, comparable to the results we achieved on SEWD. This was unexpected, because the inter-annotator agreement was, on average, substantially lower on EWD than on SEWD, which also suggested a lower performance in the automatic classification task. These results debilitate the concern that our approach might not be suitable for long turns and shows that even larger contributions can be reliably tagged with dialog acts labels. Table 6.10 shows an overview of the classifier performance both on the undersampled, balanced dataset and the full, unbalanced dataset, while figure 6.11 compares it to the performance of the human annotation task and the same two baselines used on the SEWD corpus.

To our knowledge, none of the related work on discourse analysis of Wikipedia Talk pages performed automatic dialog act classification. However, there has been previous work on classifying speech acts in other discourse types. Kim et al. (2010a) use Support Vector Machines (SVM) and Transformation Based Learning (TBL) for the automatic assignment of five speech acts to posts taken from student online forums. They report individual F_1 -scores per label which result in a macro average of 0.59 for SVM and 0.66 for TBL. Cohen et al. (2004) classify speech acts in emails. They train five binary classifiers using several learners on 1,375 emails and report F_1 scores per speech act between 0.44 and 0.85.

Despite the larger tagset, we achieved an average F_1 -score of 0.82 on the SEWD corpus and 0.78 on the EWD corpus, which compares to the top results in the related work. In future work, the performance on the dialog act classification task can be further improved by leveraging the higher flexibility of the DKPro TC framework, a flexible generalization of the

| Label | Unbalanced | Balanced |
|----------------------|------------|----------|
| CRITCOMPL | .44 | .75 |
| CRITACC | .51 | .69 |
| CRITLANG | .49 | .77 |
| CRITSUIT | .47 | .76 |
| CRITSTRUCT | .60 | .85 |
| CRITAUTH | .38 | .72 |
| ACTF | .66 | .77 |
| ACTP | .71 | .87 |
| REQEDIT | .53 | .77 |
| ATTPOS | .69 | .81 |
| ATTNEG | .53 | .78 |
| Macro average | .55 | .78 |
| Micro average | .56 | .78 |

Table 6.10: F_1 -Scores for the classifiers trained on the EWD-corpus obtained with 10-fold cross-validation both on a balanced and the full, unbalanced dataset.

FlawFinder system described in chapter 5, in order to integrate a larger number of features and utilize the parameter optimization capabilities of the framework to tune the hyperparameters of the classifiers trained. Also, the use of sequence classification algorithms might be able to make better use of the sequential nature of the discourse than our solution, in which we incorporate features from the previous and the next turn into the representation of each turn to be classified.

6.7 Application Scenario

In order to illustrate the applicability of dialog act classification in Wikipedia Talk page discussions for the information quality management process, we now discuss an application scenario in which the dialog act classifiers are used in a practical setting.

As we have established before, the global discussion activities in the English Wikipedia are on the rise and constitute the main outlet for work coordination in the open Wikipedia community (Schneider et al., 2010; Stvilia et al., 2008). At the same time, the unstructured nature of the Talk pages causes the entry barrier for new community members while exacerbating the navigation through the growing discussion archives.

An enhancement of the discussion subsystem in the MediaWiki software could drastically improve the user experience and increase the productivity of the community. This has often been suggested both by community members and researchers and activities in this area are ongoing¹³⁰. Moving from simple Wiki pages as a medium for communication to a dedicated, structured discussion system requires substantial investments on the software

¹³⁰http://www.mediawiki.org/wiki/Flow_Portal

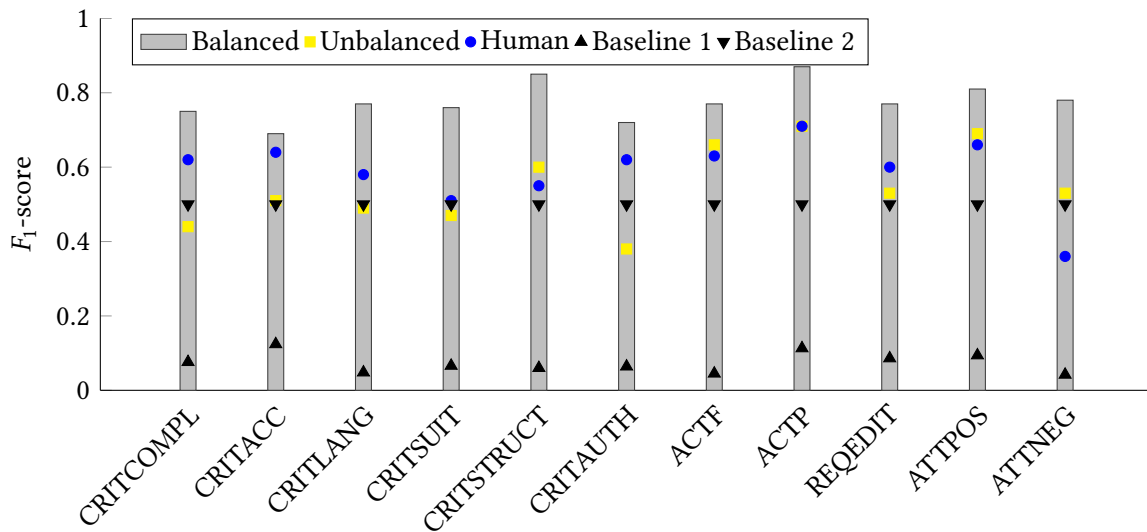


Figure 6.11: F_1 -Scores for the classifiers trained on the balanced and unbalanced dataset, the human performance and two random baselines on the EWD corpus. Baseline 1 assigns labels according to their frequency distribution in the unbalanced dataset, while baseline 2 assigns labels at random on the balanced dataset.

side. Furthermore, content created before the change to a new discussion system will not be available in a structured form in the new system and the information thus might become unavailable in the long run.

Our dialog segmentation algorithm as well as the dialog act classifiers can be used as a more lightweight solution by integrating them either as a MediaWiki plugin or an external third-party script (see the Wikimedia Labs discussed in chapter 3.6.2) on top of the existing system. This way, it is possible to provide a more structured representation of the discourse with added meta information that can be used for filtering and searching through the discussion archives. Rather than providing only a pure chronological listing of all past discussions, an augmented user interface allows users to select the aspects of the discourse they are interested in.

6.8 Chapter Summary

In this chapter, we discussed how the information on Wikipedia article Talk pages can be leveraged for information quality management purposes by automatically tagging the contributions with dialog act labels capturing the coordination efforts regarding article improvement.

We first presented an approach to reliably segment the unstructured discourse into individual discussion threads and user turns with the help of the revision history. Thereby,

we are able to retrieve additional meta information for each turn, such as the identity of its author, which is not possible by relying on the optional user signatures alone as it has been done in related work.

We furthermore present two corpora extracted from the Simple English Wikipedia and the English Wikipedia, which we manually annotated with a novel annotation scheme aimed at reflecting information quality management activities.

After a detailed analysis of these corpora and the manual annotations, we employed the data for training machine learning classifiers in order to automatically label unseen turns with the dialog act labels from our scheme. We achieved an average cross-validated performance of $F_1 = 0.82$ on the smaller and simpler SEWD corpus while we reach an average performance of $F_1 = 0.78$ on the larger EWD corpus.

These results suggest that the classifiers can be utilized in practical systems aimed at supporting the information quality management process in Wikipedia. For instance, the dialog act information can be used to filter the unstructured discussions in order to identify open issues. Furthermore, together with the segmentation algorithm presented before, it can be used to improve access to the old discussions contained in the ever growing discussion archives.

CHAPTER 7

Summary and Conclusions

“We have finished the job, what shall we do with the tools?”

— Haile Selassie

Information quality management in open collaborative environments is a complex yet vital task. In this work, we have approached the topic from the natural language processing perspective with the overarching question how language technology can help to improve quality management processes in large communities of open content production such as Wikipedia and presented two use cases – quality flaw detection in Wikipedia articles and dialog act analysis of article Talk pages. In this chapter, we give a brief summary of each chapter and provide an outlook on future research directions.

7.1 Summary

Collaboration We discussed the foundations of open collaboration in chapter 2 and introduced the main characteristics of collaborative writing, how it differs from individual writing and how open online collaboration adds an additional level of complexity to the writing task. We furthermore provided a brief description of successful systems for collaborative online writing with a particular focus on wiki technology.

Wikipedia In chapter 3, we then narrowed our focus on Wikipedia as one of the most successful online platforms for open content production and discussed its main structures and properties, its community and different approaches to process the large amounts of data the resource contains. We established that the policies governing Wikipedia and shaping its content are collaboratively defined and change over time. While large parts of these policies are shared across the different language versions, each edition has an individual take on the philosophy, which leads to a different culture in each Wikipedia.

Even though Wikipedia contains an almost incomprehensibly large set of rules and guidelines, the basic principles can be boiled down to the five pillars of Wikipedia which build the foundation for a soft security system. A unique characteristic of Wikipedia is the revision history that is kept for every page and which allows keeping track of every change ever made to the encyclopedia. At the same time, the revision history is the reason for the large amount of data Wikipedia sums up to, which makes it difficult to process as a whole.

User communication is mainly performed on the Talk pages, an unstructured discussion space in dedicated namespaces. Article Talk pages are used to coordinate the article development and discuss the future fate of an article. User talk pages, on the other hand, are used as the main means of communication between the users. There are different ways to access Wikipedia ranging from direct access of the live databases via a web API over manual processing of downloadable XML dumps to dedicated, database-driven programming interfaces. The best solution depends on the applications' need for data currency and speed and is always a compromise.

While the main reason for Wikipedia's success is its policy that everyone can contribute, the same policy also constitutes the greatest challenge. In order to establish Wikipedia as a trustworthy and comprehensive reference work with a quality level equal to edited encyclopedias, Wikipedia needs a quality management process that can cope with the almost anarchic culture that Wikipedia is based on. Taking into account the unparalleled size of the larger Wikipedia editions, a satisfactory solution can only be reached with computational assistance.

Information Quality In chapter 4, we discussed the concept of information quality and its application for information quality management. We have established that information quality, in the broadest sense, is a measure of the "fitness for use" of an information entity in a given application scenario. While it is not possible to define a single universal model of information quality, the models differ in how far they have been adapted to a particular application, medium or user group. The notion of text quality refers to an information quality model for textually represented information which particularly takes the writing quality of a text into account. In order to construct an information quality model for Wikipedia articles, we reviewed the existing mechanism for information quality management in Wikipedia to gain an overview how the concept of quality is interpreted in this community. Based on the widely accepted generic IQ model by [Wang and Strong](#), we then described an article quality model with 23 dimensions in four layers that particularly includes writing quality as a major component. The role of this model is to provide a means of orientation with respect to the aspects of quality that can be assessed with our proposed methods and also show the gaps that remain.

Quality Flaw Detection Chapter 5 contains one of the main contributions of this work. We presented an approach to automatically identify quality flaws in Wikipedia articles by means of cleanup template prediction. While cleanup templates are good proxies for quality flaws and thus a viable resource for compiling quality flaw corpora as training data for machine learning classifiers, we found that many templates exhibit a topic bias that negatively influences the classifier performance and even biases manual analyses.

We found that certain templates exhibit a topical preference, i.e. they tend to occur in articles about particular subjects, or even show a topical restriction, i.e. the templates exclusively occur in articles about particular topics. This fact has to be taken into account when sampling the data for quality flaw corpora in order to avoid a topic bias that influences both any data analyses and machine learning classifiers trained on this data.

We therefore introduced an approach to extract reliable positive and negative training instances from the article revision history which factors out the topics bias and improves the overall data quality.

We furthermore presented a corpus of articles with neutrality and style flaws that has been sampled with this technique. Our machine learning experiments on this corpus show that the reliable classifiers tend to exhibit a lower cross-validated performance than the classifiers trained on the biased datasets but the scores more closely resemble their actual performance in the wild.

We closed the chapter by describing an approach for mining quality flaw corrections from the revision history. This method can both be used to create a new parallel corpus of flawed and unflawed language as well as for identifying quality flaws within articles rather than just identifying flawed articles.

Dialog Analysis The second major contribution of this work was introduced in chapter 6, where we discussed how the content of Wikipedia article Talk pages can be leveraged for information quality management purposes by automatically tagging the user contributions with dialog act labels capturing the coordination efforts regarding article improvement.

We first presented an approach to reliably segment the unstructured discourse into individual discussion threads and user turns with the help of the revision history. Thereby, we are able to retrieve additional meta information for each turn, such as the identity of its author, which is not possible by relying on the optional user signatures alone as it has been done in related work.

We furthermore present two corpora extracted from the Simple English Wikipedia and the English Wikipedia, which we manually annotated with a novel annotation scheme aimed at reflecting information quality management activities.

After a detailed analysis of these corpora and the manual annotations, we employed the data for training machine learning classifiers in order to automatically label unseen turns with the dialog act labels from our scheme. We achieve an average cross-validated

performance of $F_1 = 0.82$ on the smaller and simpler SEWD corpus while we reach an average performance of $F_1 = 0.78$ on the larger and more complex EWD corpus.

These results suggest that the classifiers can be utilized in practical systems aimed at supporting the information quality management process in Wikipedia. For instance, the dialog act information can be used to filter the unstructured discussions in order to identify open issues. Furthermore, together with the segmentation algorithm presented before, it can be used to improve access to old discussions contained in the ever growing discussion archives.

7.2 Future Research Directions

In this final section, we identify the limitations of the current work and how they can be addressed in future work. We furthermore identify future research directions that can build upon the work presented in this thesis.

Connections between Article Discussions and Article Revisions. In this work, we have discussed the applicability of dialog analysis of Wikipedia Talk pages for improving the information quality management process, in particular the work coordination aspects. However, the Talk pages are also an invaluable resource for gaining deeper insights into the collaborative writing process. The side-by-side development of the articles on the one hand and the associated meta discussions on the other hand, which refer to evolution of the article, are unparalleled information sources for analyzing the interaction between text reception and production by the so-called prosumers. Prosumers are members of collaborative content production communities who switch between the roles of consumers and producers of information. The interaction between these two processes has a unique impact on the resulting content that has so far not been researched in detail due to a lack of sufficient data. Bringing the dialog analysis together with related work on processing the article revision history therefore promises a leap forward in understanding open collaboration in writing. A first study on this topic has recently been carried out by [Daxenberger and Gurevych \(2014\)](#), who utilize the Talk page segmentation approach presented in chapter 6 for creating their dataset. The authors achieve an accuracy of 0.86 in automatically identifying corresponding pairs of article edits and discussion turns on the article Talk page with the help of a machine learning classifier. In the future, such a system could be combined with the dialog act tagset proposed in this thesis in order to obtain a more comprehensive understanding of the relation between work coordination and edit activities.

Dialog Segmentation Accuracy and Speed. To the best of our knowledge, the dialog segmentation algorithm introduced in chapter 6.4.1 was the first approach to go beyond mere markup parsing and use the revision history of the Talk pages as an additional information

source for the segmentation process. This enabled us to reflect phenomena such as discontinuous turns or inserted replies, which is not possible with a markup based segmenter. However, as the discussion pages get older and thus the revision history grows, the speed of the segmentation drastically breaks down. While this is not a big issue for batch processing, it impedes the applicability in a real time setting, for instance in order to improve the Talk page presentation and organization on the client side without having to alter the system setup on the wiki server.

Dialog Act Sequences and Co-Occurrences. In this thesis, we aimed to account for the sequential aspects of the dialog by incorporating the text of the previous and next turn as additional features for any given turn. However, this approach cannot make use of previous classifier decisions when labeling a given turn with dialog acts. When classifying a turn, it might not only be useful to look at the text of the previous turn, but also at the labels the previous turn received. This can be achieved with a sequence classification approach, such as Conditional Random Fields or Hidden Markov Chains, which can make better use of the inter-turn dependencies. The same rationale also applies to dependencies between labels within a given turn. That is, future work should also take into account label co-occurrences in the classification task.

With a deeper incorporation of vertical (inter-turn) and horizontal (intra-turn) dialog act patterns, it will be possible to develop models that can predict the future, i.e. the dialog acts of the upcoming turns in a discussion. This, in turn, leads to interesting applications such as predicting whether a current thread is already resolved or demands further attention of the community.

Granularity of Annotation Units. We have found that the granularity of the annotation units used in the dialog act labeling task and the quality flaw recognition have a strong impact on the reliability of the training data.

In the dialog act labeling task, we have used turns as basic annotation units. Each turn is manually labeled with multiple dialog act labels. As we have discussed in chapter 6, this is a viable approach for shorter turns, but turn-level annotation is subject to a low inter-annotator agreement if long turns are involved. By extension, these turns also cause problems in the automatic classification tasks, since they introduce too much noise for a precise classification. Future work should therefore consider to use utterances as annotation units instead of labeling whole turns. This will add the additional complexity of utterance segmentation which is a research topic of its own. While some sentences can contain multiple utterances, a single utterance could also span multiple sentences. It has to be evaluated if the added complexity is justified by the possible gains in data reliability.

In the quality flaw prediction task, we have used cleanup labels assigned to whole Wikipedia articles by Wikipedia users. Thus, the annotation units are whole documents. The

annotation study in chapter 5 has shown that a human annotator can have problems in reliably identifying the presence of a single flaw in a whole document. The ability to decide the presence or absence of a flaw in an article strongly depends on the nature of the flaw and how it is represented in the article. While it is fairly easy to identify a bad article structure from looking at the article as a whole, it is more difficult to reliably identify grammatical mistakes in a longer text. Thus, structural flaws might be well represented by article-scope flaw markers while fine-grained flaws, such as language errors, should be marked directly in the text as inline- or section-scope flaws. We have already presented a method to perform the quality flaw prediction task on the sentence level in chapter 5.6 and encourage future work to extend on this approach. It has to be decided for each flaw individually on which level of granularity the given flaw should be handled.

Domain Adaptation of Quality Flaw Classifiers. This work is largely focused on Wikipedia as its main subject of analysis. While Wikipedia is one of the largest online communities for open, collaborative content production, it is not the only platform that could benefit from technology assisted information quality management. A yet unsolved question is how we can extrapolate the knowledge gained in the context of Wikipedia for improving information quality assessment processes in other collaborative platforms. In a first step, the approaches presented in this work could be transferred to platforms based on a similar technology, i.e. the MediaWiki software. In a further step, the models learned on Wikipedia can be adapted to other resources. For example, quality flaw detection could not only be carried out on Wikipedia articles but on any arbitrary texts such as online news articles e.g. in order to detect neutrality issues or biased language. However, it is safe to assume that not all of the community defined quality flaw labels will generalize equally well. Furthermore, Wikipedia specific features have to be avoided. With an appropriate domain adaptation technique, the Wikipedia quality flaw data could be bootstrapped to be used outside of the wiki context.

Integration in the Quality Management Process. The focus of this work was to provide the theoretical foundations for improving information quality management with the help of natural language processing. However, putting theory to practical use often is a complex task on its own. We have already sketched in chapter 6.7 how the dialog act classification system may be integrated in Wikipedia as a user script hosted on the Wikimedia Labs platform. The quality flaw classifiers could furthermore be used to automatically identify quality problems to be reviewed by experienced Wikipedia users. This would reduce the manual labor necessary in the review process for featured and good articles and might eventually lead to an increase of articles marked as excellent content.

Appendix

A Open Source Software

In this section, we give an overview of the open source software that has been developed in the course of this thesis or that is based on work presented in this thesis. Since open source projects are joint efforts of several developers, the descriptions in this chapter indicate the own contribution to each project and how it relates to the work presented in this thesis.

A.1 Wikipedia Revision Toolkit

The Wikipedia revision history is a valuable resource for NLP and has been used for a wide variety of applications such as spelling error detection, text simplification, text summarization or paraphrasing (Ferschke et al., 2013). Even though article revisions are available from the official Wikipedia revision dumps, accessing this information on a large scale is still a computationally intensive and thus complex task. This is due to two main problems. First, the revision dump contains all revisions as full text. Whenever a single character is changed in an article, the whole article is stored again in full. This results in a massive amount of data which requires bulk processing on powerful hardware and does not easily allow structured access to arbitrary content. Second, without an efficient API for accessing article revisions on a large scale, any research endeavor has to reinvent the wheel whenever information from the revision history is needed.

In order to tackle these two problems, we have developed the *RevisionMachine* as part of the Wikipedia Revision Toolkit (WRT)¹³¹. First, we describe our solution to the storage

¹³¹Beside the *RevisionMachine*, the WRT also contains the *TimeMachine* which re-creates arbitrary earlier states of Wikipedia from a single revision dump (Ferschke et al., 2011). The software is open source and available under <http://jwpl.googlecode.com>. It is a joint effort of several developers under the lead of Oliver Ferschke and Torsten Zesch. The individual contributions are recorded in the public SVN history on Google Code and visualized on <http://www.ohloh.net/p/jwpl>. The algorithms used in the *RevisionMachine* are based on preliminary work by Kulesa (2008)

problem. Second, we present several use cases of the RevisionMachine, and show how its API simplifies experimental setups.

A.1.1 Revision Storage

As each revision of a Wikipedia article stores the full article text, the revision history obviously contains a lot of redundant data. The RevisionMachine makes use of this fact and utilizes a dedicated storage format which stores a revision only by means of the changes that have been made to the previous revision. For this purpose, we have tested existing diff libraries, like Javaxdelta¹³² or java-diff¹³³, which calculate the differences between two texts. However, both their runtime and the size of the resulting output was not feasible for the given size of the data. Therefore, we have developed our own diff algorithm, which is based on a *longest common substring search* and constitutes the foundation for our revision storage format.

The processing of two subsequent revisions can be divided into four steps:

- First, the RevisionMachine searches for all common substrings with a user-defined minimal length.
- Then, the revisions are divided into blocks of equal length. Corresponding blocks of both revisions are then compared. If a block is contained in one of the common substrings, it can be marked as *unchanged*. Otherwise, we have to categorize the kind of change that occurred in this block. We differentiate between five possible actions: Insert, Delete, Replace, Cut and Paste¹³⁴. This information is stored in each block and is later on used to encode the revision.
- In the next step, the current revision is represented by means of a sequence of actions performed on the previous revision.

For example, in the adjacent revision pair

r_1 : This is the very first sentence!

r_2 : This is the second sentence

r_2 can be encoded as

```
REPLACE 12 10 'second'
```

```
DELETE 31 1
```

- Finally, the string representation of this action sequence is compressed and stored in the database.

With this approach, we achieve to reduce the demand for disk space of an English Wikipedia dump from June 15, 2010 containing all article revisions from 5,470 GB to only 96 GB, i.e.

¹³²<http://javaxdelta.sourceforge.net>

¹³³<http://www.incava.org/projects/java/java-diff>

¹³⁴Cut and Paste operations always occur pairwise. In addition to the other operations, they can make use of an additional temporary storage register to save the text that is being moved.

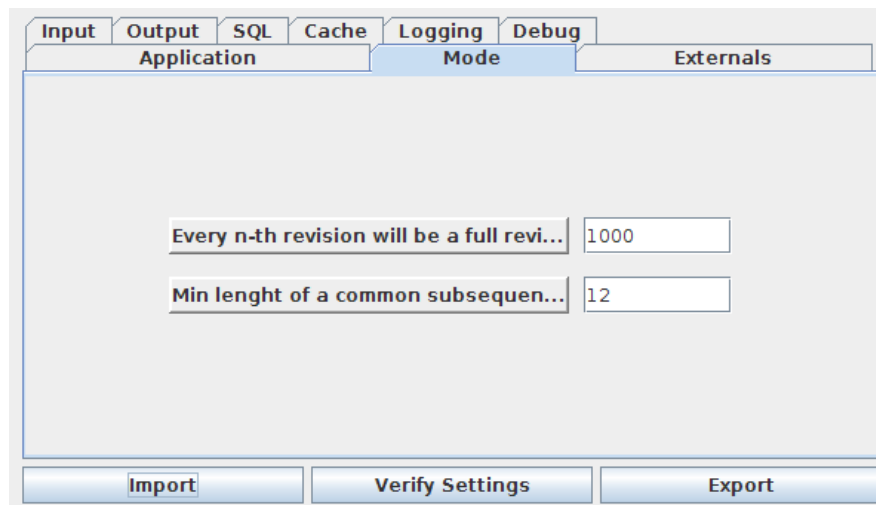


Figure A.1: Configuration GUI for the RevisionMachine

by 98%, while maintaining direct access to any data record, which is a key advantage over compressing the dump as a whole with a standard compression algorithm. The converted and compressed data records are stored in a MySQL database, which provides sophisticated indexing mechanisms for high-performance access to the data.

Obviously, storing only the changes instead of the full text of each revision trades in speed for space. Accessing a certain revision now requires to reconstruct the text of the revision from a list of changes. As articles often have several thousand revisions, this might take too long. Thus, in order to speed up the recovery of the revision text, every n -th revision is stored as a full revision. A low value of n decreases the time needed to access a certain revision, but increases the demand for storage space. We have found $n = 1000$ to yield a good trade-off. If hard disk space is no limiting factor, the parameter can be set to 1 or another small number to avoid the compression of the revisions and maximize the performance. This parameter, among a few other possibilities to fine-tune the process, can be set in a graphical user interface provided with the RevisionMachine. (see figure A.1).

A.1.2 Revision Access

After the converted revisions have been stored in the revision database, the database can either be used stand-alone in combination with additional data extracted from the Wikipedia dump by the JWPL (Zesch et al., 2008). The latter option makes it possible to combine the possibilities of the RevisionMachine with other components like the JWPL parser for the MediaWiki syntax.

In order to set up the RevisionMachine, it is only necessary to provide the configuration details for the database connection (see listing A.1). Upon first access, the database user has to have write permission on the database, as indexes have to be created. For later use, read

```

//Set up database connection
DatabaseConfiguration db = new DatabaseConfiguration();
db.setDatabase("dbname");
db.setHost("hostname");
db.setUser("username");
db.setPassword("pwd");
db.setLanguage(Language.english);
//Create API objects
Wikipedia wiki = WikiConnectionUtils.getWikipediaConnection(db);
RevisionIterator revIt = new RevisionIterator(db);
RevisionApi revApi = new RevisionApi(db);

```

Listing A.1: Setting up the RevisionMachine

```

//Iterate over all revisions of all articles
while (revIt.hasNext()) {
    Revision rev = revIt.next();
    rev.getTimestamp();
    rev.getArticleID();
    //process revision ...
}

```

Listing A.2: Iteration over all revisions of all articles

permission is sufficient. Access to the RevisionMachine is achieved via two API objects. The *RevisionIterator* allows to iterate over all revisions in Wikipedia. The *RevisionAPI* grants access to the revisions of individual articles. In addition to that, the *Wikipedia* object provides access to JWPL functionalities.

In the following, we describe three use cases of the RevisionMachine API, which demonstrate how it is easily integrated into experimental setups.

Processing all article revisions in Wikipedia The first use case focuses on the utilization of the complete set of article revisions in a Wikipedia snapshot. Listing A.2 shows how to iterate over all revisions. Thereby, the iterator ensures that successive revisions always correspond to adjacent revisions of a single article in chronological order. The start of a new article can easily be detected by checking the timestamp and the article id. This approach is especially useful for applications in statistical natural language processing, where large amounts of training data are a vital asset.

Processing revisions of individual articles The second use case shows how the RevisionMachine can be used to access the edit history of a specific article. The example in listing A.3 illustrates how all revisions for the article *Automobile* can be retrieved by first performing a page query with the JWPL API and then retrieving all revision timestamps for this page, which can finally be used to access the revision objects.

Accessing the meta data of a revision The third use case illustrates the access to the meta data of individual revisions. The meta data includes the name or IP of the contributor, the additional user comment for the revision and a flag that identifies a revision as minor or


```

//Get article with title "Automobile"
Page article = wiki.getPage("Automobile");
int id = article.getPageId();
//Get all revisions for the article
Collection<Timestamp> revisionTimeStamps = revApi.getRevisionTimeStamps(id);
for (Timestamp t:revisionTimeStamps) {
    Revision rev = revApi.getRevision(id, t);
    //process revision ...
}

```

Listing A.3: Accessing the revisions of a specific article

```

//Meta data provided by the RevisionAPI
StringBuffer s = new StringBuffer();
s.append("The article has "+revApi.getNumberOfRevisions(pageId)+" revisions\n");
s.append("It has "+revApi.getNumberOfUniqueContributors(pageId)+" unique contributors\n");
s.append(revApi.getNumberOfUniqueContributors(pageId, true)+ " are registered users\n");
//Meta data provided by the Revision object
s.append((rev.isMinor()?"Minor":"Major")+ " revision by: "+rev.getContributorID());
s.append("\nComment: "+rev.getComment());

```

Listing A.4: Accessing the meta data of a revision

major. Listing A.4 shows how the number of edits and unique contributors can be used to indicate the level of edit activity for an article.

A.2 DKPro Text Classification Framework

The DKPro Text Classification Framework (DKPro TC)¹³⁵ is an open source text classification system that emerged from an enhancement and generalization of the FlawFinder system that was described in chapter 5.4.

For the quality flaw prediction task, we required a system for exploring a wide range of machine learning algorithms while allowing to automatically optimize the hyperparameters for each algorithm, the configuration of the preprocessing and, above all, providing easily extensible feature extraction capabilities. We already described in chapter 5.4.1 how we designed the FlawFinder system to fulfill all of the above requirements for the particular task of flaw detection.

DKPro TC takes the FlawFinder approach to the next level by scaling to generic supervised learning problems involving textual data. The main goal of DKPro TC is to move the focus away from the mere technical aspects of machine learning experiments and rather stress the importance of higher level design decisions and the development of an expressive feature set for the task at hand. Therefore, DKPro TC automates as many aspects of the

¹³⁵The software is open source and available under <http://dkpro-tc.googlecode.com>. It is a joint effort of several developers under the lead of Johannes Daxenberger, Oliver Ferschke and Torsten Zesch and based on the FlawFinder system developed by Oliver Ferschke in the course of this thesis. The individual contributions are recorded in the public SVN history on Google Code and visualized on <http://www.ohloh.net/p/dkpro-tc>







| |  Single-label |  Multi-label |  Regression |
|--|--|---|--|
|  Document Mode | · Spam Detection · Sentiment Detection | · Text Categorization · Keyphrase Assignment | · Text Readability |
|  Unit Mode | · Word Sense Disambiguation · Sentiment Aspect Detection | · Dialog Act Tagging | · Word Difficulty |
|  Pair Mode | · Paraphrase Identification · Textual Entailment | · Relation Extraction | · Text Similarity |

Table A.1: Supervised learning scenarios supported by DKPro TC with exemplary NLP applications

experiment workflow as possible while still letting the researcher control and monitor any aspect of the process.

At the time of writing, DKPro TC supports three different experiment modes

- In *document mode*, each input document is treated as an individual entity to be classified, e.g. an email classified as spam or ham.
- In *unit mode*, each input document contains several units to be classified. It is usually not possible to split the document into separate documents, because the context of each unit needs to be preserved, e.g. in word sense disambiguation with Lesk (Lesk, 1986).
- The *pair mode* is intended for problems which require a pair of texts as input, e.g. a pair of sentences to be classified as paraphrase or non-paraphrase.

Each mode can either be employed to perform a binary classification task, a multi-label classification task or a regression problem. Table A.1 gives an overview of exemplary machine learning tasks that can be solved with the individual combinations of experiment modes and learning problems.

The overall concept of DKPro TC is similar to the original FlawFinder system but has been refined and generalized in several aspects. The system architecture can be separated into the following six components which we will describe briefly in the remainder of this section.

Workflow Engine Like FlawFinder, DKPro TC uses the DKPro Lab (Eckart de Castilho and Gurevych, 2011) as a runtime environment, which allows to define configurable, task-based experiment workflows. Most of the modules described in the remainder of this section are implemented as Lab-Tasks which each contain a single UIMA pipeline. All tasks are wired together in the experiment definition in order to setup the overall experiment workflow.

Each task can furthermore be parameterized in order to control the execution of the inner NLP pipeline. This way, different settings for the UIMA processing components can be employed thus enabling the researcher to identify an optimal configuration of the experiment. The DKPro Lab thereby makes sure that intermediate output is not unnecessarily recalculated if the parameters for the given task did not change since the last execution.

In order to shield the user from the complexity of the task definitions and task wiring, several standard application scenarios, such as machine learning with cross validation or train/test evaluation, are already included in the framework and can be used out of the box.

Reading Input Data In order to make a dataset available to the DKPro TC framework, a UIMA reader has to be created that converts the dataset into the UIMA Common Analysis Structure (CAS). Depending on the experiment mode, the DKPro TC framework furthermore requires the user to define the gold standard labels for each classification unit in an *outcome* annotation which will be used for training and evaluation by the framework.

Preprocessing Once the dataset has been imported into DKPro TC, standard UIMA components can be used to preprocess the data according to the requirements of the downstream feature extractors. DKPro Core¹³⁶, for example, provides a large collection of general purpose NLP components that can be used for this purpose.

Feature Extraction Defining an expressive feature set is one of the main aspects of machine learning experiments. Feature extractors in DKPro TC are UIMA analysis engines that can make use of all the annotations created by the preprocessing components. The framework defines interfaces for the different available experiment modes which govern the behavior of the extractors. That is, while a document feature extractor extracts features from a whole CAS, unit feature extractors extract features only from a particular span of text. Document pair extractors furthermore extract features from pairs of documents. All features are stored in a global feature store which allows to make use of the extracted features independent from the downstream machine learning algorithm or toolkit.

Supervised Learning DKPro TC does not provide own implementations of machine learning algorithms but rather contains interfaces to established machine learning toolkits. At the time of writing, DKPro TC provides full support for the algorithms in the Weka Machine learning toolkit (Hall et al., 2009) while providing basic support for the classification algorithms in the Mallet toolkit (McCallum, 2002). Multi label classification experiments furthermore make use of the Mulan library (Tsoumakas et al., 2010). DKPro TC takes care of preparing the data to fit the requirements of the chosen machine learning software

¹³⁶<http://dkpro-core-asl.googlecode.com>

Evaluation and Reporting Depending on the experiment mode, DKPro TC provides an overview of the precision, recall and F_1 -scores achieved with each experiment configuration. It furthermore provides the option to record confusion matrices, list the actual predictions assigned to each document by the classifiers and give an overview of the feature rankings if a feature selection algorithm had been employed. The user is furthermore free to attach additional report modules to the experiment which can then record arbitrary additional information.

B Annotation Guidelines

In the following sections, we reproduce the annotation guidelines provided to the annotators of the SEWD and EWD corpora. The guidelines have been reformatted and shortened, where appropriate. In addition to these documents, the annotators received additional instructions and training on demand.

B.1 Annotation guidelines for the SEWD corpus

In contrast to the EWD corpus, the annotation of the SEWD corpus was carried out by two annotators who were mainly trained orally without providing an exhaustive annotator's manual. The annotators were furthermore supervised during a training period that preceded the annotation task. In this training period, a small, set of Talk pages was annotated. Afterwards, the annotations were discussed with the instructor and the annotators were asked to justify all of their decisions. The Talk pages annotated in the pre-study have not been included in the SEWD corpus. The annotation process is further described in chapter 6. Apart from the oral instructions, the annotators received the annotation scheme (see table 6.4) along with the following short instructions:

CM

Some information, statement or utterance is *not present* in the article but *should be present*.

CW

Factual errors. Some information, statement or utterance should be corrected or rephrased in order to be correct.

CU

Some information, statement or utterance is *present* in the article but should *not be present*, because it is unsuitable, unnecessary, obsolete, or too detailed.

CS

Concerns the inner structure of the article or the position of the article within the wider framework of Wikipedia. Also, merging or splitting of an article falls into this category.

CL

Unsuitable language or style, unclear formulation and any need for rephrasing in order to express the facts correctly.

COBJ

Lack of neutrality (NPOV)

CO

Any kind of criticism not covered by the categories above including fuzzy criticism (“The article/section is odd”).

PSR

- Could anybody do *X*?
- Please do *X*
- I would say somebody should *X*
- ”It should be neutral”
- ”The following should be clear”
- The section must be changed.

PREF

This class does not apply to citations or reference material included in the user contribution. Generally, it is not applied when material is referenced to support the own statement. It applies to an action of referencing or pointing to some external or internal subject matter, e.g.

- Please see *X*.
- Please look at the section *Y* in the article
- As I have stated in a previous discussion

PFC

Commits to an action in the future, e.g.

- “I will change *X*.”
- “Changing *X*.”
- “Moving *X*”

PPC

Report of performed action, e.g.

- “Done”
- “Fixed”

IP

Covers any information providing acts, such as answers, replies, elaborations, statements, announcements, quotes and comments.

IS

Any information seeking acts, such as questions. Note that rhetorical questions rarely seek for information. Thus, they should just be labeled with IP. Also, requests or suggestions alone do not always seek for information.

IC

Correcting an already established fact by providing the corrected fact. Contributions marked with this label are usually also marked with IP.

B.2 Annotation guidelines for the EWD corpus

The annotation scheme used in this study was designed to reflect the ways Wikipedia users coordinate article improvement. Your task as an annotator is to identify contributions that point out faults or a lack of quality in the discussed article, offer solutions to the identified problems, and announce actions towards improving the article. At the same time, the attitude of one participant to another on an interpersonal level is recorded.

The corpus consists of a selection of Wikipedia Talk pages taken from the English Wikipedia from April 6, 2011. Each Talk page has been segmented into discussions (i.e. the individual topics discussed on a talk page) and turns (i.e. the individual user contributions within a discussion). In order to be selected for the corpus, a Talk page must have more than one discussion and its size must be between 1,000 and 40,000 characters¹³⁷. The Talk pages are selected according to the cleanup templates that occur in the Talk page or in the article associated with the Talk page. The same number of articles from each category – *distinguished article*, *flawed article* and *neutral article* – is selected for the corpus.

B.2.1 General Guidelines

The annotation scheme has not been designed to cover all possible aspects of human conversation. It particularly focuses on the aspects described in the introduction. Consequently,

¹³⁷The articles are categorized in six size-classes: 1,000-7,500, 7,501-14,000, 14,001-25,000, 25,001-27,000, 27,001-33,500, 33,501-40,000. From each class, the same number of articles is selected

it might not be possible to label every contribution in a discussion. This is not a problem, as we are only interested in the contributions about article improvement.

Discussions are labeled first on the discussion level, i.e. the level of an individual discussion-topic in the discussion page (cf. section B.2.2) and then on the turn level, i.e. the level of the individual user contributions (cf. section B.2.3).

Segmentation Errors

In some cases, discussions are not segmented correctly, e.g. contributions of more than one user are treated as a single contribution. In those cases, the whole discussion can be marked as *rejected* by assigning the ERROR label. As a consequence, the whole discussion is rejected for further in the experiment (cf. section B.2.2).

Discontinuous Contributions

In some cases, contributions can be discontinuous, i.e. they may contain (correctly segmented) inserted contributions by other authors. When selecting such a discontinuous contribution, all parts belonging to the turn will be highlighted while leaving out the inserted contributions. The inserted contributions can be selected and annotated separately. This should not be confused with segmentation errors (i.e. a selection that highlight the contributions of more than one user at once)

Surface Structure vs. Intention

Do not be influenced too much by the surface structure of the text, i.e. do not give too much attention to the individual sentence types (question, statement). What matters is the content and the intention of each contribution. A “question” like “*Shouldn’t the invention of the transistor be given a lot more attention?*” is not just a request for information, but (also) criticism regarding a lack of detail and a suggestion to expand the information.

Certainty and Uncertainty

Certainty and uncertainty are not covered by the annotation scheme, so “*I thought that FACT_X*” or “*It might be the case that FACT_X*” are treated the same way as “*FACT_X*”.

The Role of the Topic Title

When labeling the first contribution in a discussion, you should also include the topic title in your analysis. In most cases, the discussion title has been written by the first contributor and contains additional information which might even be necessary to interpret the first turn. The title can be seen as part of the first contribution.

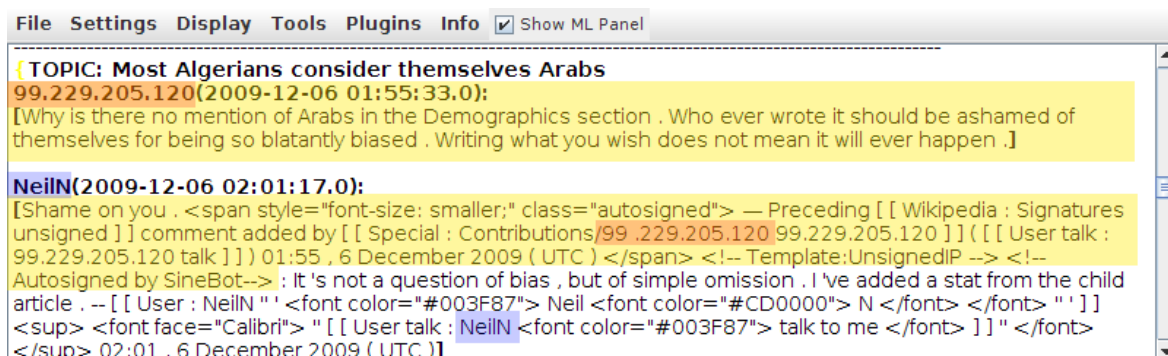


Figure B.2: Example for a segmentation error displayed in the MMAX2 annotation tools

Repetition

If a user repeats a statement that has already been already labeled e.g. as some kind of criticism in an earlier contribution, it should be labeled again as such. For an example, see section B.3. In that example, turn 3 paraphrases/repeats the criticism from Turn 1 (“*GDP is wrong and he/she knows it*”) Consequently, it has to be labeled the same way as the criticism in Turn 1. The same is true for the other label categories besides criticism.

B.2.2 Discussion Level

ERROR: Segmentation Errors

If a discussion was not segmented correctly, it might be problematic to annotate the user contributions. In that case, the ERROR label should be assigned. The discussion is then rejected for further use in this annotation experiment. No further annotation has to be performed on this discussion. A strong indicator for a segmentation error would be the presence of a user signature of some user Y WITHIN the contribution of some user X. (Because the signature suggests a contribution boundary that was missed by the parser). However, if there are signatures of X inside a contribution of X, it does not pose a problem. This is most like due to an aggregation of several subsequent contributions by the same author into a single turn. Wrong author attribution (i.e. the case that the correct author for a contribution could not be found) does not qualify for this label.

REFOBJ: Reference Object

The “Reference Object” category defines the main focus of a discussion. This category is the only non-binary one. Only one of the three possible labels can be chosen for an individual discussion.

PART: Article Part The focus of the discussion is an individual aspect or part of the article, e.g. the discussion of a particular defect, lack of quality, missing fact

WHOLE: Whole Article The focus of the discussion is the “article as a whole” and not a single detail of the article or an individual fact.

Examples

- Article protection
- Discussion page protection
- Article vandalism
- User bans
- Article status (featured, good, stub)

META: Meta Discussion The discussion is completely detached from the article (Off-Topic) or refers to resources outside of Wikipedia.

B.2.3 Turn Level

Article Criticism

All labels in this section refer to a contributions mentioning a lack of quality in one of six categories. They do not only apply to contributions explicitly stating criticism (e.g. “*FACT_X is missing*”), but also to contribution implicitly stating subjects of improvement, for example

- suggestions for improvement: “*FACT_X should be added*”
- implicit suggestions for improvement: “*shouldn’t FACT_X be added?*”)
- requests for improvement: “*ARTICLE_X states that 1+1=3. Please correct this.*” - this is criticism regarding ”Accuracy and Correctness“, cf. CRITACC)

CRITCOMPL: Incompleteness or Lack of Detail The main content of the article is incomplete and/or parts of the article are not detailed enough. Possible reasons for choosing this label

- Lack of detail
- Missing facts
- Missing images
- Other missing content
- Suggestions for content that should be added
- ”X needs clarification“

Not to be confused with

- Missing references: ⇒ CRITAUTH
- Missing links: ⇒ CRITSTRUCT
- Missing templates: ⇒ CRITSTRUCT
- Missing categories: ⇒ CRITSTRUCT
- Incorrect or obsolete content: ⇒ CRITACC

Note: This label does not apply to missing structural elements or references, only to the main text body of the article.

CRITACC: Lack of Accuracy, Correctness and Neutrality The article or part of the article is inaccurate, incorrect or biased/not neutral. Possible reasons for choosing this label

- Wrong facts
- Erroneous content
- Content not up to date (real world has changed - content has to be updated)
- Inaccurate description
- NPOV (non-neutral point of view)
- Biased content
- Article is not objective
- "Poor quality" (if no specific information is given about the reason)
- Wrong terminology used (in combination with CRITLANG)
- Passage not understandable ("What does X mean") if there's no indication that it's due to CRITLANG-issues or missing explanation (CRITCOMPL)

Not to be confused with

- Unclear or fuzzy formulation: ⇒ CRITLANG (in some cases, both labels can be assigned to one contribution)
- Missing facts or images ⇒ CRITCOMPL

Note: In the actual discussion, this label might occur when people rephrase text from the article. In this case, it must be determined whether the paraphrase was really made to **correct incorrect or inaccurate information** (then this is the right label). If the paraphrase was made to improve the language in order to make the statement more understandable or less ambiguous, *Language and Style* (CRITLANG) is the correct label.

CRITLANG: Deficiencies in Language and Style The article or part of the article contains bad language or style. Possible reasons for choosing this label

- Typing slips, language errors
- Language too complex
- Low readability, obscure language
- Too much use of foreign language
- Unclear formulation (the content might still be accurate and correct)
- Ambiguous formulation
- Inconsistent use of vocabulary: using different vocabulary for the same concept within the article or category
- Text is incohesive
- Text does not flow well
- Terminological issues (wrong terminology used): Depending on the context, it can be combined with CRITACC

Not to be confused with

- Incorrect or inaccurate statements: ⇒ CRITACC (in some cases, both labels can be assigned to one contribution)
- Structural problems, formal issues(citation style), formatting issues: ⇒ CRITSTRUCT

Note: In the actual discussion, this label might occur when people rephrase text from the article. In this case, it must be determined whether the paraphrase was really made to **improve the language in order to make the statement more understandable or less ambiguous** (then this is the right label). If the paraphrase was made to correct incorrect or inaccurate information, *Accuracy and Correctness* (CRITACC) is the correct label.

CRITSUIT: Unsuitability The article or part of the article contains unsuitable content. Possible reasons for choosing this label

- Content redundancy
- Content irrelevant or outside the scope of the article
- Content too detailed
- Content of an image used in the article is unsuitable
- Quality of an image used in the article is unsuitable
- Media with unsuitable licence / non-free image
- Content that violates copyright (e.g. *copy and paste-text*)
- Vandalism

Not to be confused with

- Missing licence information: ⇒ CRITAUTH

CRITSTRUCT: Deficiencies in Structure, Organization and Visual Appearance The article or part of the article has a bad structure or is visually not appealing or it is not correctly placed and connected within the broader framework of Wikipedia. Possible reasons for choosing this label

- Poor organization/structure of the article
- Content/sections should be rearranged/reordered
- Headings should be renamed
- Nonconformity to the suggested style guides
- Inconsistent usage of structures and styles: content is structured differently than in similar articles
- Missing/inadequate/too many TEMPLATES
- Missing/inadequate/too many TAGS
- Missing/inadequate/too many CATEGORIES
- Missing/inadequate/too many LINKS
- Too many redlinks
- Dead (external) links
- Article should be split into several articles

- Article should be merged with other article
- Cosmetic problems with the typesetting

Not to be confused with

- Incorrect language, typos: ⇒ CRITLANG

CRITAUTH: Lack of Authority The article lacks authority and verifiability or has (media) licencing issues . Possible reasons for choosing this label

- Lack of supporting sources
- Plagiarism
- Lack of academic scrutiny of the sources (sources are given, but they are not reliable)
- Known bias of the sources
- Lack of references to original sources
- Lack of accessibility of original sources
- Claims or details in the article cannot be verified
- Contains uncited content
- Lack of proper licencing information (for media)

Not to be confused with

- NPOV (non-neutral point of view) ⇒ CRITACC
- Media with unsuitable (non-free) licence ⇒ CRITSUIT

Self Commitment

ACTF: Commitment to Future Action This label relates to announcements of future article-related actions. It is chosen if the author

- announces future commitment (I will do this)
- offers future commitment (I could do this. I can do this. If nobody else does it, I might do it)

ACTP: Report of Past Action This label relates to reports of already performed article-related actions. It is chosen if the author

- reports an action (e.g. "I already fixed the error in the article")
- uses a template like fixed or done

Requests

Request-labels do not cover all possible requests, like e.g. the request of a discussion contributor that someone else should fix an error in the article. This "implicit" request is already contained in the criticism labels and is not encoded again if the request is made explicitly. The request labels only cover request that fit one of the following three classes.

REQEDIT: Request for Article Edit This label is chosen if the author requests that the article should be edited (e.g. to fix an error that was identified by a contribution with a CRIT* label. **Note:** This label is used if discussion contributors ask the community to edit the article. If the user announces to do the editing him/herself, the ACTF label should be assigned. If the user reports an already performed action, the ACTP label should be assigned. If the requested action is an admin or maintenance action and not a simple article edit, the REQMAINT label should be assigned.

REQMAINT: Request for Admin/Maintenance Action This label is chosen if the author

- specifically requests an admin to protect/semiprotect the article
- specifically requests an admin to remove article protection
- specifically requests an admin or reviewer to review the article for promotion/demotion (e.g. to featured / good status)
- specifically requests an admin to join two articles
- specifically requests an admin to split an article in two (or more) articles
- specifically requests an admin to move the article to a different namespace
- specifically requests other maintenance actions

Note: This label specifically addresses maintenance actions that cannot be performed by normal users. Simple article edits or adding the article to a category are not covered by this label (use REQEDIT for that). This label also includes the request for an article review by an admin or reviewer to evaluate if the article should be promoted/demoted to featured/good status!

Interpersonal

The interpersonal categories are only to be used if the attitude towards another participant of the discussion is made explicit. It should only be used to characterize the attitude of an author towards another user or their contributions and/or whether they agree or disagree with other contributions. The labels are polar - states between positive or negative do not exist. If such a case occurs, it should not get an Interpersonal label. If an author shows positive attitude towards one user and negative attitude towards another user, both labels can be assigned. **The attitudes towards people not taking part in the discussion is not covered by the annotation scheme and should not be labeled with *Interpersonal* labels.**

ATTPOS: Positive Attitude / Support / Agreement This label is chosen if the author

- supports/agrees with the contribution/opinion/idea of another author
- confirms the contribution of another author
- accepts the contribution/opinion of another author (“I agree”, “You’re right”)
- compliments another author (“Good work”)
- praises another contribution or author

- shows gratitude (“Thanks”)
- shows appreciation (“I like the idea that...”)

ATTNEG: Negative Attitude / Reject / Disagreement This label is chosen if the author

- rejects or objects to the contribution/opinion/idea of another author (either by explicitly expressing an opinion or by taking the counter-position)
- disagrees with the contribution/opinion/idea of another author (“I disagree”, “You are wrong”)
- threatens another author (“If you don’t stop, I’ll report you”)
- dislikes another author or their contribution (“I am not fond of this way of thinking”)
- blames another author (“You messed up the article”)

B.2.4 Example

Figure B.3 shows a very short example discussion loaded into the annotation tool MMAX2. The discussion only consists of 3 contributions. The following subsections show, how this discussion should be annotated:



Figure B.3: Example discussion about the article *Algeria* in MMAX2

Discussion Level

The main focus of this discussion is a wrong figure in the article about *Algeria*. Thus, the reference object of the discussion is PART. It can be argued that, as the discussion develops, the focus changes towards blocking a “Troll” from the article, which would demand the

label **WHOLE**, but the anchor of the whole discussion is the incorrect GDP value, so the final label should be **PART**.

There are no segmentation errors in this discussion, so the **ERROR** label must not be assigned (i.e. its value should remain *false*).

Turn 1: 2010-12-01

The user criticizes a wrong figure in the article (**CRITACC**). The request of the user that someone should change the value is not modeled in the annotation scheme.

Turn 2: 2010-12-30

The user agrees with the contribution of the user in *Turn 1* (**ATTPOS**). He reports that he has corrected the error in the article (**ACTP**). He further requests that the article be protected, which is a maintenance action (**REQUMAINT**).

Turn 3: 2011-01-12

The user in *Turn 3* again request maintenance action, i.e. to block a user (**REQUMAINT**). He further repeats the criticism mentioned in *Turn 1*, (**CRITACC**). The negative attitude towards the “Troll” is not modeled in the annotation scheme, because the “Troll” takes not part in the discussion.

B.2.5 Cases with unclear label assignments

Reference object cannot clearly be chosen

Some discussions contain contributions with a clear **PART** focus and, at the same time, contributions with a clear **WHOLE** focus. If the number of contributions that indicate one particular reference object is much larger than the other, use the the one with the bigger support. If this is not the case, use the label that you would choose when only reading the first contribution of the discussion.

CRITACC or CRITLANG?

Sometimes, the boundaries between **CRITACC** and **CRITLANG** become fuzzy. For example, in the contribution

How can the Tomb of the Unknown Soldier in Paris have 'inspired' the Tomb of the Unknown Soldier in Westminster Abbey? They were both done at the same time.

the lack of accuracy and correctness is caused by unclear formulation. In this case, both labels can be assigned. This is also the case when the author discusses whether it is correct to express something in a certain way, e.g. in this contribution

Is it right to say 'the' tomb of the Unknown soldier, shouldn't that be the the French tomb of the Unknown soldier or the tomb of the Unknown soldier in France ?

In most cases, however, you can decide for one of the two labels.

C Cleanup Templates in the English Wikipedia

This section lists all cleanup templates of the English Wikipedia as of 16 July 2012 according to <http://en.wikipedia.org/wiki/WP:TC> excluding writing variants and synonymous templates listed in the “see also” sections of the template description pages. The functional groups are based on the categorization in the original template listing, but have been adapted where appropriate. Wherever possible, we assigned to each category the corresponding quality dimensions defined in chapter 4. Since these assignments are done per category and not per label, a category could contain outliers that do not fit the dimensions assigned to the category.

General

cleanup, cleanup AfD, cleanup-remainder, cleanup-rewrite, cleanup-articletitle

Copy Edit (→ Grammaticality and Spelling, Word choice, Understandability, Conventions)

copy edit, copy edit-section

Subject Specific

CIA, cleanup Congress Bio, cleanup-book, cleanup-chartable, cleanup-comics, cleanup-IPA, cleanup-school, cleanup-university, game cleanup, game guide, hadith authenticity, local, metricate, toLCleanup, USRD-wrongdir

Fiction

all plot, book-fiction, fiction, in-universe, plot, dubious conversion, need-IPA

Style of Writing (→ Tone, Conventions, Word Choice)

abbreviations, db-spam, buzzword, cleanup-tense, crystal, debate, editorial, essay-like, howto, inappropriate person, like resume, news release, db-spam, news, release section, obituary, pro and con list, repetition, review, story, technical, tone, travel guide, over-quotation, capitalization

Structure and format (→ Structure)

cleanup-reorganize, importance-section, section-diffuse, sections, spacing, lead missing, format footnotes, sub-sections

Amount of information (→ Amount of Information)

condense, duplication, too many see alsos, very long, lead too long, lead too short

Unwanted Content (→ Amount of Information, Value added, Connectivity)

cleanup-spam, Cleanup Red Link, close paraphrasing, cypypaste, criticism section, external links, further reading cleanup, in popular culture, MOS, non-free, NOT, overlinked, schedule, trivia, contact information, spam link, off-topic

Context and detail (→ Amount of information, Complexity)

context, generalize, generalize-section, over detailed, specific

Expand and add (→ Completeness)

cleanup-biography, cleanup-weighted, expand section, formula missing descriptions, ISBN, kmposts, mileposts, Lacking overview, missing information, biblio

Time-sensitive (→ Currency, Volatility)

out of date, recently revised, time-context, update, update after, clarify timeframe

Contradiction and Confusing (→ Understandability, Accuracy)

confusing, contradict, contradict-other, contradict-other-multiple, incoherent, incoherent-topic, misleading, unclear date, contradiction-inline, expand, acronym, inconsistent, vague,

Importance and Notability (→ Value added)

notability, puffery

Accuracy (→ Accuracy)

disputed, disputed-section, dubious, clarify, bad unit conversions, bad summary, lead rewrite, inadequate lead, expert-subject, Expert-talk, expert-verify

Neutrality (→ Neutrality)

advert, cherry picked, coat rack, COI, geographical imbalance, globalize, peacock, POV, Neutrality, POV-check, POV-lead, POV-section, POV-title, recentism, unbalanced, undue, weasel, peacock-inline, weasel-inline, editorializing, lopsided, POV-statement

Reliability, Reputation and Trustworthiness (→ Reputation)

verify credibility, unreliable medical source, unreliable sources, vague, verify source, syn, original research, or, self-published, dubious, elucidate, examples, failed verification, self-published inline

Verifiability (→ Verifiability)

BLP IMDb refimprove, BLP sources, BLP sources section, BLP unsourced, citation style, citations broken, citations missing, cite check, cleanup-link rot, ibid, , better source, medref, more footnotes, no footnotes, one source, page numbers improve, page numbers needed, primary sources, refimprove, ref improve section, religious text primary, symbolism, third-party, unreferenced, unreferenced section, film IMDb refimprove, attribution needed, by whom, chronology, citation needed, citation broken, citation needed, citation needed (lead), cite quote, clarify, copyvio link, dead link, disambiguation needed, full, medical citation needed, nonspecific, page needed, quantify, registration required, request quotation, season needed, specify, subscription required, third-party-inline, volume needed, when, where, which?, who, whom?, whosequote, why?, year needed, find sources, find sources 3, search

Categories (→ Categorization)

cat improve, category relevant?, category unsourced, , recategorize, uncategorized, uncategorized stub

Images (→ Illustration)

cleanup-gallery, cleanup-images, image requested, reqdiagram, reqmap, reqscreenshot, too many photos

Lists

cleanup-laundry, create-list, disputed-list, list fact, example farm, fictionrefs, in popular culture, list to table, MOSLOW, prose

WikiTech (→ Connectivity)

cleanup-HTML, dead end, disambiguation cleanup, disambiguation, incoming links, more-specific-links, orphan, wikify, shadowsCommons, prod

Infobox

infobox requested, newinfobox, ship infobox request, single infobox request, cleanup-infobox

Merge

Afd-merge from, afd-merged-from, Afd-merge to, merge, merge from, merge to, merged-from, merged-to, merging, cleanup-combine,

Move

move header, move to userspace, movenotice, moveoptions, convert to SVG and copy to Wikimedia Commons, copy to Meta, copy to Wikibooks, copy to Wikibooks, Cookbook,

softredirect, copy to Wikimedia Commons, copy to Wikiquote, copy to Wikisource, copy to Wikiversity, now Commons

Split (→ Structure)

cleanup split, split, split-apart, split dab, split section, split sections

Translations and Language issues (→ Grammaticality and Spelling, Understandability, Word choice)

cleanup-translation, expand Spanish, not English, not English-inline, rough translation, translated page, translatePassage, translation WIP, TWCleanup

Completeness (→ Completeness)

Stub¹³⁸

¹³⁸In addition to the generic stub template, topic specific stub-templates are available and more commonly used. A list can be found under <http://en.wikipedia.org/wiki/WP:STUBSHORT>

List of Tables

| | | |
|------|--|-----|
| 3.1 | Number of pages per namespace for the five largest Wikipedias | 23 |
| 3.2 | Wikipedia namespaces and functional Talk page classes | 35 |
| 3.3 | Tools and services for accessing Wikipedia | 39 |
| 4.1 | Number of articles per WikiProject quality level | 55 |
| 5.1 | Agreement of human annotator with gold standard. The corpus for this small study consist of 20 articles per flaw, half of which are flawed. | 71 |
| 5.2 | Flaw definitions and numbers of training and test instances per flaw in the CLEF corpus | 75 |
| 5.3 | NSTYLE corpus of neutrality and style flaws | 81 |
| 5.4 | Cosine similarity scores between the category frequency vectors of the flawed article sets and the respective random or reliable negatives | 84 |
| 5.5 | Positive and negative instances per flaw in NSTYLE | 87 |
| 5.6 | Feature sets used in the experiments on the CLEF and NSTYLE corpora | 92 |
| 5.7 | Overview of the feature utility scores on the CLEF corpus | 96 |
| 5.8 | Average F_1 -scores over all flaws on NSTYLE-RELP using NSTYLE-ALL features | 97 |
| 5.9 | Comparison of classifier performance on CLEF across participants | 100 |
| 5.10 | Overview of classification errors per flaw on CLEF. | 101 |
| 5.11 | F_β scores for the 10-fold cross validation of the SVMs with RBF kernel on all datasets using NSTYLE-NGRAM features | 103 |
| 5.12 | Sample sentence pairs from the uncertainty corpus | 105 |
| 6.1 | Page-level features proposed by Kittur et al. (2007) | 116 |
| 6.2 | Overview of authority claims in Wikipedia discussions | 118 |
| 6.3 | Descriptive statistics for the Simple English Wikipedia and the English Wikipedia | 130 |
| 6.4 | Annotation scheme for the SEWD corpus | 134 |

| | | |
|------|---|-----|
| 6.5 | Annotation scheme for the EWD corpus | 135 |
| 6.6 | Label frequencies and inter-annotator agreement for the SEWD corpus. . . . | 140 |
| 6.7 | Label frequencies and inter-annotator agreement for the EWD corpus | 141 |
| 6.8 | Distribution of dialog acts according to topic scope | 145 |
| 6.9 | Classifier performance on the SEWD corpus | 149 |
| 6.10 | Classifier performance on the EWD corpus | 151 |
| A.1 | Supervised learning scenarios supported by DKPro TC with exemplary NLP applications | 166 |

List of Figures

| | | |
|------|--|-----|
| 1.1 | NLP-assisted information quality management in Wikipedia | 3 |
| 2.1 | Collaborative writing strategies | 13 |
| 3.1 | Main page of the English Wikipedia on 2 Sept 2013 | 20 |
| 3.2 | The origin of Wikipedia policies | 21 |
| 3.3 | Structure of an article | 27 |
| 3.4 | Revisions of the article “Natural Language Processing” | 32 |
| 3.5 | Structure of a Talk page | 34 |
| 3.6 | Examples for user signatures on Talk pages | 36 |
| 3.7 | DTD for the MediaWiki export format | 38 |
| 4.1 | Hierarchical information quality model after Wang and Strong | 48 |
| 4.2 | The rating interface of the Article Feedback Tool | 57 |
| 4.3 | Proposed model for article quality in Wikipedia | 59 |
| 5.1 | Cleanup templates with page-, section-, and inline scope | 69 |
| 5.2 | Dimensional coverage of the article quality model by cleanup templates | 73 |
| 5.3 | Concept of one-class classification | 76 |
| 5.4 | Concept of PU learning | 78 |
| 5.5 | Sampling of negative instances for quality flaw detection | 79 |
| 5.6 | Distributed extraction of reliable negative instances using Hadoop | 82 |
| 5.7 | Descriptive statistics for the NSTYLE corpus | 86 |
| 5.8 | High-level system architecture of the FlawFinder | 88 |
| 5.9 | Classifier performance on CLEF in terms of precision, recall and F_1 -score. | 98 |
| 5.10 | Classifier performance on the NSTYLE corpus | 102 |
| 6.1 | Creation and utilization of the Wikipedia Talk page corpora | 121 |
| 6.2 | In-text replies on a Talk page for the article <i>Monaro Highway</i> | 122 |

| | | |
|------|---|-----|
| 6.3 | Identification of paragraph creation points with forward checking. | 124 |
| 6.4 | Identification of paragraph creation points with backward checking. | 125 |
| 6.5 | Selection criteria for Talk pages in the EWD corpus | 131 |
| 6.6 | Mapping between the SEWD and EWD annotation schemes | 136 |
| 6.7 | Turn Annotation in MMAX2 | 138 |
| 6.8 | Topic Annotation in MMAX2 | 138 |
| 6.9 | Expert support system for creating the EWD gold standard | 139 |
| 6.10 | Classifier performance on the SEWD corpus | 150 |
| 6.11 | Classifier performance on the EWD corpus | 152 |
| A.1 | Configuration GUI for the RevisionMachine | 163 |
| B.2 | Example for a segmentation error | 172 |
| B.3 | Example discussion about the article <i>Algeria</i> in MMAX2 | 178 |

Bibliography

Khalid Al Khatib, Hinrich Schütze, and Cathleen Kantner: ‘Automatic Detection of Point of View Differences in Wikipedia’, in: *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pp. 33–50, Mumbai, India, 2012. (Cited on page 22)

Maik Anderka: *Analyzing and Predicting Quality Flaws in User-generated Content : The Case of Wikipedia*, Phd dissertation, Bauhaus Universität Weimar, 2013,
Online: http://www.uni-weimar.de/medien/webis/publications/papers/anderka_2013.pdf.
(Cited on page 85)

Maik Anderka and Benno Stein: ‘Overview of the 1st International Competition on Quality Flaw Prediction in Wikipedia’, in: *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers*, Rome, Italy, 2012. (Cited on pages 74, 75, 99, 100 and 104)

Maik Anderka, Benno Stein, and Nedim Lipka: ‘Predicting Quality Flaws in User-generated Content: The Case of Wikipedia’, in: *35th International ACM Conference on Research and Development in Information Retrieval*, pp. 981–990, Portland, OR, USA, 2012. (Cited on pages 66, 70, 76, 77 and 97)

Ofer Arazy and Rick Kopak: ‘On the measurability of information quality’, *Journal of the American Society for Information Science and Technology* 62 (1): 89–99, January 2011.
(Cited on pages 50 and 71)

Ron Artstein and Massimo Poesio: ‘Inter-Coder Agreement for Computational Linguistics’, *Computational Linguistics* 34 (4): 555–596, 2008. (Cited on page 140)

John Longshaw Austin: *How to Do Things with Words*, Clarendon Press, Cambridge, UK, 1962. (Cited on page 112)

Phoebe Ayers, Charles Matthews, and Ben Yates: *How Wikipedia Works: And how You Can be a Part of it*, No Starch Press Series, No Starch Press, 2008. (Cited on pages 20, 26 and 30)

- Daniel Bär, Nicolai Erbs, Torsten Zesch, and Iryna Gurevych: ‘Wikulu: An Extensible Architecture for Integrating Natural Language Processing Techniques with Wikis’, in: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. System Demonstrations*, pp. 74–79, 2011. (Cited on page 16)
- Regina Barzilay and Mirella Lapata: ‘Modeling local coherence: An entity-based approach’, in: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 141–148, 2005. (Cited on page 53)
- Robert-Alain de Beaugrande and Wolfgang Ulrich Dressler: *Introduction to Text Linguistics*, Longman linguistics library, Longman, 1981. (Cited on pages 53 and 60)
- Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf: ‘Annotating Social Acts: Authority Claims and Alignment Moves in Wikipedia Talk Pages’, in: *Proceedings of the Workshop on Language in Social Media*, pp. 48–57, Portland, OR, USA, 2011. (Cited on pages 117, 118 and 143)
- Christian Bizer: *Quality-Driven Information Filtering- In the Context of Web-Based Information Systems*, VDM Verlag, Saarbrücken, Germany, July 2007. (Cited on page 49)
- Carl-Hugo Björnsson: *Läsbarhet: Lesbarkeit durch Lix. (Aus dem Schwedischen)*, (Pedagogiskt Utvecklingsarbete vid Stockholms Skolor. 6.), Liber, 1968. (Cited on page 93)
- Julian Brooke and Graeme Hirst: ‘Native Language Detection with ‘Cheap’ Learner Corpora’, in: *Proceedings of Learner Corpus Research 2011*, Louvain-la-Neuve, Belgium, 2011. (Cited on page 78)
- Michel Buffa: ‘Intranet wikis’, in: *Proceedings of the IntraWebs Workshop 2006 at the 15th International World Wide Web Conference*, Edinburgh, Scotland, UK, 2006. (Cited on page 16)
- Karl Bühler: *Sprachtheorie: Die Darstellungsfunktion der Sprache*, Verlag von Gustav Fischer in Jena, 1934. (Cited on page 112)
- Harry Bunt and William Black: ‘The ABC of computational pragmatics’, in Harry Bunt and William Black (Eds.): *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics*, pp. 1–46, John Benjamins Publishing Co., 2000. (Cited on page 113)
- Jean Carletta: ‘Assessing agreement on classification tasks: The kappa statistic’, *Computational Linguistics* 22 (2): 249–254, 1996. (Cited on pages 72 and 139)
- Tamitha Carpenter and Emi Fujioka: ‘The Role and Identification of Dialog Acts in Online Chat’, in: *Workshops at the 25th AAAI Conference on Artificial Intelligence*, San Francisco, CA, USA, 2011. (Cited on page 114)

- Imogen Casebourne, Chris Davies, Michelle Fernandes, and Naomi Norman: ‘Assessing the accuracy and quality of Wikipedia entries compared to popular online encyclopaedias: A comparative preliminary study across disciplines in English, Spanish and Arabic.’, *Technical report*, Epic, Brighton, 2012. (Cited on page 1)
- William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell: ‘Learning to Classify Email into “Speech Acts”’, in: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 309–316, Barcelona, Spain, 2004. (Cited on pages 114 and 150)
- Meri Coleman and T. L. Liau: ‘A computer readability formula designed for machine scoring.’, *Journal of Applied Psychology* 60 (2): 283, 1975. (Cited on pages 52 and 93)
- Mark G. Core and James F. Allen: ‘Coding Dialogues with the DAMSL Annotation Scheme’, in: *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, November, 1997. (Cited on page 113)
- Holly Crawford: ‘Encyclopedias’, in Richard E. Bopp and Linda C. Smith (Eds.): *Reference and Information Services: An Introduction*, pp. 433–459, Libraries Unlimited, Eaglewood, CO, USA, 3rd edition, 2001. (Cited on page 58)
- Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg: ‘Echoes of power: Language effects and power differences in social interaction’, in: *Proceedings of the 21st International World Wide Web Conference (WWW 2012)*, pp. 699–708, Lyon, France, 2012. (Cited on pages 118 and 122)
- Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch: ‘DKPro TC: A Java-based Framework for Supervised Learning Experiments on Textual Data’, in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, pp. 61–66, Baltimore, MD, USA, 2014. (Cited on page 87)
- Johannes Daxenberger and Iryna Gurevych: ‘Automatically Detecting Corresponding Edit-Turn-Pairs in Wikipedia’, in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Short Papers*, p. (to appear), Baltimore, MD, USA, 2014. (Cited on page 158)
- Han De Vries, Marc N Elliott, David E Kanouse, and Stephanie S. Teleki: ‘Using Pooled Kappa to Summarize Interrater Agreement across Many Items’, *Field Methods* 20 (3): 272–282, March 2008. (Cited on page 140)
- Hannes Dohrn and Dirk Riehle: ‘Design and implementation of the Sweble Wikitext parser’, in: *Proceedings of the International Symposium on Wikis and Open Collaboration (WikiSym ’11)*, pp. 72–81, Mountain View, CA, USA, 2011a. (Cited on page 89)
- Hannes Dohrn and Dirk Riehle: ‘Wom: An object model for wikitext.’, *Technical report*, University of Erlangen, Erlangen, Germany, 2011b. (Cited on page 89)

- Richard Eckart de Castilho and Iryna Gurevych: ‘A Lightweight Framework for Reproducible Parameter Sweeping in Information Retrieval’, in: *Proceedings of the Workshop on Data Infrastructures for Supporting Information Retrieval Evaluation*, pp. 7–10, Glasgow, UK, 2011. (Cited on pages [88](#) and [166](#))
- Martin Eppler: *Managing Information Quality - Increasing the Value of Information in Knowledge-intensive Products*, Springer Berlin / Heidelberg, 1st edition, 2003. (Cited on page [50](#))
- Martin Eppler and Dörte Wittig: ‘Conceptualizing Information Quality: A Review of Information Quality Frameworks from the Last Ten Years’, in: *Proceedings of the 2000 Conference on Information Quality*, 2000. (Cited on pages [46](#), [47](#) and [48](#))
- Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas: ‘The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text’, in: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL ’10: Shared Task, pp. 1–12, Uppsala, Sweden, 2010. (Cited on page [106](#))
- Edgardo Ferretti, Donato Hernández Fusilier, Rafael Guzmán-Cabrera, Manuel Montes-y Gómez, Marcelo Errecalde, Paolo Rosso, and Manuel Montes y Gómez: ‘On the Use of PU Learning for Quality Flaw Prediction in Wikipedia’, in: *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers*, Rome, Italy, 2012. (Cited on pages [77](#), [78](#), [97](#) and [100](#))
- David Ferrucci and Adam Lally: ‘UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment’, *Natural Language Engineering* 10 (3-4): 327–348, 2004. (Cited on pages [87](#) and [146](#))
- Oliver Ferschke, Johannes Daxenberger, and Iryna Gurevych: ‘A Survey of NLP Methods and Resources for Analyzing the Collaborative Writing Process in Wikipedia’, in Iryna Gurevych and Jungi Kim (Eds.): *The People’s Web Meets NLP: Collaboratively Constructed Language Resources*, Theory and Applications of Natural Language Processing, chapter 5, Springer, 2013. (Cited on pages [33](#) and [161](#))
- Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar: ‘Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages’, in: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 777–786, Avignon, France, 2012a. (Cited on pages [34](#), [93](#) and [110](#))
- Oliver Ferschke, Iryna Gurevych, and Marc Rittberger: ‘FlawFinder: A Modular System for Predicting Quality Flaws in Wikipedia’, in: *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers*, Rome, Italy, 2012b. (Cited on page [100](#))
- Oliver Ferschke, Torsten Zesch, and Iryna Gurevych: ‘Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia’s Edit History’, in: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

- System Demonstrations*, pp. 97–102, Portland, OR, USA, 2011. (Cited on pages 40, 93 and 161)
- Jenny Rose Finkel, Trond Grenager, and Christopher Manning: ‘Incorporating non-local information into information extraction systems by Gibbs sampling’, in: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 363–370, Association for Computational Linguistics, Morristown, NJ, USA, June 2005. (Cited on page 93)
- Aidan Finn and Nicholas Kushmerick: ‘Learning to classify documents according to genre’, *Journal of the American Society for Information Science and Technology* 57 (11): 1506–1518, 2006. (Cited on page 78)
- Lucie Flekova, Oliver Ferschke, and Iryna Gurevych: ‘What Makes a Good Biography? Multidimensional Quality Analysis Based on Wikipedia Article Feedback Data’, in: *Proceedings of the 23rd International World Wide Web Conference (WWW 2014)*, pp. 855–866, Seoul, Korea, April 2014. (Cited on page 56)
- Rudolf Flesch: ‘A new readability yardstick.’, *The Journal of applied psychology* 32 (3): 221, 1948. (Cited on pages 52 and 93)
- Andrea Forte, Aniket Kittur, Vanessa Larco, Haiyi Zhu, Amy Bruckman, and Robert E. Kraut: ‘Coordination and Beyond: Social Functions of Groups in Open Content Production’, in: *Proceedings of the 2012 ACM Conference on Computer Supported Cooperative Work*, Seattle, WA, USA, 2012. (Cited on page 10)
- Andrea Forte and Cliff Lampe: ‘Defining, Understanding, and Supporting Open Collaboration: Lessons From the Literature’, *American Behavioral Scientist* 2013. (Cited on pages 10 and 11)
- Akinori Fujino, Hideki Isozaki, and Jun Suzuki: ‘Multi-label Text Categorization with Model Combination based on F1-score Maximization’, in: *Proceedings of the Third International Joint Conference on Natural Language Processing*, pp. 823–828, Hyderabad, India, 2008. (Cited on pages 71 and 146)
- Jolene Galegher and Robert E. Kraut: ‘Computer-mediated communication for intellectual teamwork: a field experiment in group writing’, *Information Systems Research* 5 (2): 110–138, September 1994. (Cited on page 12)
- Mouzhi Ge: *Information quality assessment and effects on inventory decision-making*, Phd dissertation, Dublin City University, 2009. (Cited on page 50)
- Jim Giles: ‘Internet encyclopaedias go head to head’, *Nature* 438 (7070): 900, 2005. (Cited on page 1)
- Ruediger Glott, Philipp Schmidt, and Rishab Ghosh: ‘Wikipedia Survey – Overview of Results’, *Technical report*, United Nations University, Maastricht, 2010. (Cited on page 30)

- Johannes Gordesch and Burkhard Dretzke: ‘Correctness in language: A formal theory’, *Journal of Quantitative Linguistics* 5 (1-2): 13–26, 1998. (Cited on page 51)
- Arthur C. Graesser, Danielle S. McNamara, and Jonna M. Kulikowich: ‘Coh-Metrix: Providing Multilevel Analyses of Text Characteristics’, *Educational Researcher* 40 (5): 223–234, June 2011. (Cited on page 53)
- Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai: ‘Coh-metrix: analysis of text on cohesion and language.’, *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc* 36 (2): 193–202, May 2004. (Cited on page 53)
- Herbert Paul Grice: ‘Logic and Conversation’, in Peter Cole and Jerry L Morgan (Eds.): *Syntax and Semantics*, Vol. 3, New York: Academic Press, 1975. (Cited on page 142)
- Barbara Grosz and Candace Sidner: ‘Attention, intentions, and the structure of discourse’, *Computational linguistics* 12 (3), 1986. (Cited on page 53)
- Barbara Grosz, Scott Weinstein, and Aravind Joshi: ‘Centering: A framework for modeling the local coherence of discourse’, *Computational linguistics* 21: 203–225, 1995. (Cited on page 53)
- Jonathan Grudin and Steven Poltrock: *Computer Supported Cooperative Work*, chapter 27, The Interaction Design Foundation, Aarhus, Denmark, 2013. (Cited on page 9)
- Robert Gunning: ‘The fog index after twenty years’, *Journal of Business Communication* 6 (2): 3–13, 1969. (Cited on pages 52 and 93)
- Iryna Gurevych, Max Mühlhäuser, Christof Müller, Jürgen Steimle, Markus Weimer, and Torsten Zesch: ‘Darmstadt Knowledge Processing Repository Based on UIMA’, in: *Proceedings of the First Workshop on Unstructured Information Management Architecture at Biannual Conference of the Society for Computational Linguistics and Language Technology*, Tübingen, Germany, 2007. (Cited on pages 89 and 93)
- Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian Witten: ‘The WEKA Data Mining Software: An Update’, *SIGKDD Explorations* 11 (1): 10–18, 2009. (Cited on pages 90, 94, 146 and 167)
- Michael A. K. Halliday and Ruqaiya Hasan: *Cohesion in English*, Longman, 1976. (Cited on pages 53, 60 and 62)
- Jingyu Han, Xiong Fu, Kejia Chen, and Chuandong Wang: ‘Web Article Quality Assessment in Multi-dimensional Space’, in: *Proceedings of the 12th International Conference on Web-age Information Management*, pp. 214–225, Springer Berlin Heidelberg, Wuhan, China, 2011a. (Cited on page 66)

- Jingyu Han, Chuandong Wang, and Dawei Jiang: ‘Probabilistic Quality Assessment Based on Article’s Revision History’, in: *Proceedings of the 22nd International Conference on Database and Expert Systems Applications*, pp. 574–588, Toulouse, France, 2011b. (Cited on page 66)
- Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, Pável Calado, Daniel Hasan Dalip, and Marcos André Gonçalves: ‘Automatic Quality Assessment of Content Created Collaboratively by Web Communities’, in: *Proceedings of the Joint International Conference on Digital Libraries*, pp. 295–304, ACM Press, Austin, TX, USA, June 2009. (Cited on page 66)
- Benjamin Mako Hill and Aaron Shaw: ‘The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation’, *PLoS ONE* 8 (6), 2013. (Cited on page 30)
- Johannes Hoffart, Torsten Zesch, and Iryna Gurevych: ‘An Architecture to Support Intelligent User Interfaces for Wikis by Means of Natural Language Processing’, in: *Proceedings of the International Symposium on Wikis and Open Collaboration (WikiSym ’09)*, Orlando, FL, USA, October 2009. (Cited on page 16)
- Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto: ‘Collaboratively built semi-structured content and Artificial Intelligence: The story so far’, *Artificial Intelligence* 194: 2–27, January 2013. (Cited on page 22)
- George Hripcsak and Adam S Rothschild: ‘Agreement, the F-Measure, and Reliability in Information Retrieval’, *Journal of the American Medical Informatics Association* 12 (3): 296–298, 2005. (Cited on page 139)
- Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong: ‘Measuring article quality in wikipedia: models and evaluation’, in: *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, pp. 243–252, Lisbon, Portugal, 2007. (Cited on page 66)
- Ian Hutchby and Robin Wooffitt: *Conversation Analysis*, John Wiley & Sons, 2nd edition, 2008. (Cited on page 113)
- Roman Jakobson: ‘Closing Statements: Linguistics and Poetics’, *Style in Language* pp. 350–377, 1960. (Cited on page 112)
- Sara Javanmardi and Cristina Lopes: ‘Statistical measure of quality in Wikipedia’, in: *Proceedings of the First Workshop on Social Media Analytics*, pp. 132–138, ACM Press, Washington DC, DC, USA, July 2010. (Cited on page 66)
- Michael T Joyce: ‘Siren Shapes: Exploratory and Constructive Hypertexts’, *Academic Computing* 3, 1988. (Cited on page 15)

- Dan Jurafsky: ‘Pragmatics and Computational Linguistics’, in Laurence R. Horn and Gregory Ward (Eds.): *Handbook of Pragmatics*, pp. 578–604, Blackwell Publishing Ltd, 2006. (Cited on pages 112 and 113)
- Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca: ‘Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual’, *Technical Report Draft 13*, University of Colorado, Institute of Cognitive Science, 1997. (Cited on page 113)
- Jihie Kim, Jia Li, and Taehwan Kim: ‘Towards Identifying Unresolved Discussions in Student Online Forums’, in: *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 84–91, Los Angeles, CA, USA, 2010a. (Cited on pages 114, 143 and 150)
- Su Nam Kim, Li Wang, and Timothy Baldwin: ‘Tagging and linking web forum posts’, in: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning, CoNLL ’10*, pp. 192–202, Stroudsburg, PA, USA, 2010b. (Cited on page 143)
- Peter Kincaid, Robert Fishburne Jr, Richard Rogers, and Brad Chissom: ‘Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.’, *Technical report*, DTIC Document, 1975. (Cited on pages 52 and 93)
- Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, H. Chi, and Ed H. Chi: ‘He Says, She Says: Conflict and Coordination in Wikipedia’, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 453–462, San Jose, CA, USA, 2007. (Cited on pages 115, 116 and 185)
- Shirlee-ann Knight and Janice Burn: ‘Developing a framework for assessing information quality on the World Wide Web’, *Informing Science Journal* 8, 2005. (Cited on page 46)
- Olavi Koistinen: ‘World’s largest study on Wikipedia: Better than its reputation’, 2013, Online: <http://www.helsinkitimes.fi/finland/finland-news/domestic/8619-world-s-largest-study-on-wikipedia-better-than-it-s-reputation.html>. (Cited on page 1)
- Moshe Koppel and Jonathan Schler: ‘Exploiting Stylistic Idiosyncrasies for Authorship Attribution’, in: *Workshop on Computational Approaches to Style Analysis and Synthesis*, pp. 69–72, 2003. (Cited on page 78)
- Klaus Krippendorff: *Content Analysis: An Introduction to Its Methodology*, Sage Publications, 1980. (Cited on page 141)
- Klaus Krippendorff: ‘Der verschwundene Bote. Metaphern und Modelle der Kommunikation’, in Klaus Merten, Siegfried J. Schmidt, and Siegfried Weischenberg (Eds.): *Die Wirklichkeit der Medien*, pp. 79–113, VS Verlag für Sozialwissenschaften, 1994. (Cited on page 111)

- Simon Kulesa: *Extracting paraphrases from the Wikipedia revision history*, Diplomarbeit, Technische Universität Darmstadt, 2008. (Cited on page 161)
- Shyong K. Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R. Musicant, Loren Terveen, and John Riedl: ‘WP:Clubhouse? An Exploration of Wikipedia’s Gender Imbalance’, in: *Proceedings of the International Symposium on Wikis and Open Collaboration (WikiSym ’11)*, Mountain View, CA, 2011. (Cited on page 30)
- J Richard Landis and Gary Koch: ‘An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers’, *Biometrics* 33 (2): 363–374, 1977. (Cited on pages 139 and 141)
- David Laniado, Riccardo Tasso, Andreas Kaltenbrunner, Politecnico Milano, and Yana Volkovich: ‘When the Wikipedians Talk : Network and Tree Structure of Wikipedia Discussion Pages’, in: *Proceedings of the 5th International Conference on Weblogs and Social Media*, pp. 177–184, Barcelona, Spain, 2011. (Cited on pages 118, 119 and 121)
- Andrew LaVallee: ‘Jimmy Wales on Wikipedia Quality and Tips for Contributors’, 2009, Online: <http://blogs.wsj.com/digits/2009/11/06/jimmy-wales-on-wikipedia-quality-and-tips-for-contributors>. (Cited on page 2)
- Yang W. Lee, Diane M. Strong, Beverly K. Kahn, and Richard Y. Wang: ‘AIMQ: a methodology for information quality assessment’, *Information & Management* 40 (2): 133–146, December 2002. (Cited on page 47)
- Michael Lesk: ‘Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone’, in: *Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC ’86*, pp. 24–26, ACM, Toronto, Canada, 1986. (Cited on page 166)
- Bo Leuf and Ward Cunningham: *The Wiki Way: Quick Collaboration on the Web*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2001. (Cited on page 16)
- Sharman Lichtenstein and Craig M Parker: ‘Wikipedia model for collective intelligence: a review of information quality’, *International Journal of Knowledge and Learning* 5 (3/4): 254–272, 2009. (Cited on page 58)
- Charles Ling and Victor Sheng: ‘Cost-Sensitive Learning’, in Claude Sammut and Geoffrey I. Webb (Eds.): *Encyclopedia of Machine Learning*, pp. 231–235, Springer US, 2010. (Cited on page 147)
- Nedim Lipka and Benno Stein: ‘Identifying featured articles in wikipedia’, in: *Proceedings of the 19th International World Wide Web Conference (WWW 2010)*, p. 1147, ACM Press, Raleigh, NC, USA, April 2010. (Cited on page 66)
- Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S. Yu: ‘Building Text Classifiers Using Positive and Unlabeled Examples’, in: *Proceedings of the Third IEEE International*

- Conference on Data Mining*, pp. 179–186, November 2003. (Cited on page 77)
- Bing Liu, W. S. Lee, P. S. Yu, and Xiaoli Li: ‘Partially supervised classification of text documents’, *Proceedings of the 19th International Conference on Machine Learning* 2002. (Cited on page 77)
- Annie Louis: *Predicting text quality: Metrics for content, organization and reader interest*, Phd, University of Pennsylvania, 2013. (Cited on pages 52, 53 and 61)
- Paul B. Lowry, Aaron Curtis, and Michelle René Lowry: ‘Building a Taxonomy and Nomenclature of Collaborative Writing to Improve Interdisciplinary Research and Practice’, *Journal of Business Communication* 41 (1): 66–99, January 2004. (Cited on pages 12, 13, 14 and 15)
- Kim Luyckx and Walter Daelemans: ‘Shallow Text Analysis and Machine Learning for Authorship Attribution’, in: *Proceedings of the Fifteenth Meeting of Computational Linguistics in the Netherlands (CLIN 2004)*, pp. 149–160, Utrecht, Netherlands, 2004. (Cited on page 78)
- Alex Marin, Bin Zhang, and Mari Ostendorf: ‘Detecting Forum Authority Claims in Online Discussions’, in: *Proceedings of the Workshop on Languages in Social Media*, pp. 39–47, Portland, OR, USA, June 2011. (Cited on page 117)
- Paolo Massa: ‘Social Networks of Wikipedia’, in: *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, pp. 221–230, Eindhoven, Netherlands, 2011. (Cited on page 119)
- Paolo Massa and Federico Scrinzi: ‘Exploring Linguistic Points of View of Wikipedia’, in: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pp. 213–214, New York, NY, USA, 2011. (Cited on page 22)
- Andrew McCallum: ‘MALLET: A Machine Learning for Language Toolkit’, 2002, Online: <http://mallet.cs.umass.edu>. (Cited on pages 90, 94 and 167)
- Philip McCarthy, Gwyneth Lewis, David Dufty, and Danielle McNamara: ‘Analyzing Writing Styles with Coh-Metrix’, in: *Proceedings of the Florida Artificial Intelligence Research Society International Conference*, pp. 764–769, Melbourne Beach, FL, USA, 2006. (Cited on page 53)
- G Harry McLaughlin: ‘SMOG grading: A new readability formula’, *Journal of Reading* 12 (8): 639–646, 1969. (Cited on pages 52 and 93)
- Christian M. Meyer: *Wiktionary - The Metalexicographic and the Natural Language Processing Perspective*, Phd dissertation, Technische Universität Darmstadt, 2013, Online: <http://tuprints.ulb.tu-darmstadt.de/3654>. (Cited on page 41)

- George K. Mikros and Eleni K. Argiri: ‘Investigating Topic Influence in Authorship Attribution’, in: *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*, Amsterdam, Netherlands, 2007. (Cited on page 78)
- David Milne and Ian Witten: ‘An Open-source Toolkit for Mining Wikipedia’, in: *Proceedings of the New Zealand Computer Science Research Student Conference*, Auckland, New Zealand, 2009. (Cited on page 40)
- Thomas Mitchell: *Machine Learning*, McGraw-Hill Education, New York, NY, USA, 1st edition, 1997. (Cited on pages 94 and 147)
- Christoph Müller and Michael Strube: ‘Multi-level annotation of linguistic data with MMAX2’, in Sabine Braun, Kurt Kohn, and Joybrato Mukherjee (Eds.): *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pp. 197–214, Peter Lang, Frankfurt a.M., DE, 2006. (Cited on page 137)
- Eugene W. Myers: ‘An O(ND) Difference Algorithm and Its Variations’, *Algorithmica* 1: 251–266, 1986. (Cited on page 105)
- Masaaki Nagata, Yumi Shibaki, and Kazuhide Yamamoto: ‘Using Goi-Taikai as an Upper Ontology to Build a Large-Scale Japanese Ontology from Wikipedia’, in: *Proceedings of the 6th Workshop on Ontologies and Lexical Resources (Ontolex 2010)*, pp. 11–18, Beijing, China, 2010. (Cited on page 25)
- Sylvie Noël and Jean-Marc Robert: ‘How the Web is used to support collaborative writing’, *Behaviour & Information Technology* 22 (4): 245–262, 2003. (Cited on page 15)
- Sylvie Noël and Jean-Marc Robert: ‘Empirical Study on Collaborative Writing: What Do Co-authors Do, Use, and Like?’, *Computer Supported Cooperative Work* 13 (1): 63–89, 2004. (Cited on page 15)
- Philip V. Ogren, Philipp G. Wetzler, and Steven Bethard: ‘ClearTK: A UIMA toolkit for statistical natural language processing’, in: *Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP workshop at Language Resources and Evaluation Conference (LREC)*, pp. 32–38, Marrakech, Morocco, 2008. (Cited on page 89)
- Felipe Ortega: *Wikipedia: A Quantitative Analysis*, Phd dissertation, Universidad Rey Juan Carlos, 2009, Online: <http://ciencia.urjc.es/handle/10115/11239>. (Cited on pages 2 and 41)
- Felipe Ortega, Jesus M. Gonzalez-Barahona, and Gregorio Robles: ‘On the Inequality of Contributions to Wikipedia’, in: *Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, HICSS ’08, pp. 304–, IEEE Computer Society, Washington, DC, USA, 2008. (Cited on page 11)

- Meghan Oxley, Jonathan T. Morgan, and Brian Hutchinson: ‘“What I Know Is...”: Establishing Credibility on Wikipedia Talk Pages’, in: *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, pp. 2–3, Gdańsk, Poland, 2010. (Cited on pages [117](#) and [118](#))
- Rebecca J. Passonneau: ‘Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation’, in: *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy, 2006. (Cited on page [141](#))
- Emily Pitler and Ani Nenkova: ‘Revisiting readability: a unified framework for predicting text quality’, in: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, number October in EMNLP ’08, pp. 186–195, Association for Computational Linguistics, Honolulu, HI, USA, 2008. (Cited on page [52](#))
- John C. Platt: ‘Fast training of support vector machines using sequential minimal optimization’, in Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola (Eds.): *Advances in Kernel Methods: Support Vector Learning*, pp. 185–208, MIT Press, 1998. (Cited on pages [97](#) and [147](#))
- Ross Quinlan: *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers Inc., 1st edition, 1992. (Cited on page [147](#))
- Laura Rassbach, Trevor Pincock, and Brian Mingus: ‘Exploring the Feasibility of Automatically Rating Online Article Quality’, in: *Proceedings of the 2007 International Wikimedia Conference (WikiMania)*, Taipei, Taiwan, 2007. (Cited on page [66](#))
- Eric S. Raymond: *The Cathedral and the Bazaar*, O’Reilly & Associates, Inc., Sebastopol, CA, USA, October 1999. (Cited on page [10](#))
- Joseph Michael Jr. Reagle: *Good Faith Collaboration: The Culture of Wikipedia (History and Foundations of Information Science)*, The MIT Press, 2010. (Cited on pages [20](#) and [30](#))
- Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky: ‘Linguistic Models for Analyzing and Detecting Biased Language’, in: *Proceedings of the 51st Annual Meeting on Association for Computational Linguistics*, pp. 1650–1659, Sofia, Bulgaria, 2013. (Cited on page [105](#))
- Scott Rettberg: ‘All together now: Collective knowledge, collective narratives, and architectures of participation’, in: *Proceeding of the 2005 Digital Arts and Culture Conference*, Copenhagen, DK, 2005. (Cited on pages [14](#) and [22](#))
- R. P. Rice and Jr. Huguley J.T.: ‘Describing collaborative forms: a profile of the team-writing process’, *IEEE Transactions on Professional Communication* 37 (3): 163–170, 1994. (Cited on page [12](#))
- Jan P. Rohweder, Gerhard Kasten, Dirk Malzahn, Andrea Piro, and Joachim Schmid: ‘Informationsqualität – Definitionen, Dimensionen und Begriffe’, in Knut Hildebrand,

- Marcus Gebauer, Holger Hinrichs, and Michael Mielke (Eds.): *Daten- und Informationsqualität*, chapter 2, pp. 25–45, Vieweg+Teubner Verlag, Wiesbaden, 2008. (Cited on pages [47](#) and [48](#))
- Jerrold Sadock: ‘Speech Acts’, in Laurence R. Horn and Gregory Ward (Eds.): *Handbook of Pragmatics*, chapter 3, pp. 53–73, Blackwell Publishing Ltd, 2006. (Cited on page [112](#))
- Markus Schaal, Barry Smyth, Roland M. Mueller, and Rutger MacLean: ‘Information quality dimensions for the social web’, in: *Proceedings of the International Conference on Management of Emergent Digital EcoSystems - MEDES '12*, p. 53, ACM Press, New York, New York, USA, October 2012. (Cited on page [48](#))
- Jodi Schneider, Alexandre Passant, and John G. Breslin: ‘A Content Analysis: How Wikipedia Talk Pages Are Used’, in: *Proceedings of the 2nd International Conference of Web Science*, pp. 1–7, Raleigh, NC, USA, 2010. (Cited on pages [110](#), [115](#) and [151](#))
- Jodi Schneider, Alexandre Passant, and John G. Breslin: ‘Understanding and Improving Wikipedia Article Discussion Spaces’, in: *Proceedings of the 2011 ACM Symposium on Applied Computing*, pp. 808–813, Taichung, Taiwan, 2011. (Cited on pages [57](#), [110](#), [115](#) and [121](#))
- John Rogers Searle: *Speech Acts*, Cambridge University Press, Cambridge, UK., 1969. (Cited on page [112](#))
- John Rogers Searle: ‘A classification of illocutionary acts’, *Language in Society* 5 (1): 1–23, 1976. (Cited on pages [112](#) and [113](#))
- Claude E Shannon: ‘A Mathematical Theory of Communication’, *Bell System Technical Journal* 27 (3): 379–423, 1948. (Cited on pages [45](#) and [46](#))
- Mike Sharples: ‘Introduction’, in Mike Sharples (Ed.): *Computer Supported Collaborative Writing*, Computer Supported Cooperative Work, pp. 1–7, Springer London, 1993. (Cited on page [12](#))
- Mike Sharples, J. Goodlet, Eevi Beck, C. Wood, S. Easterbrook, and L. Plowman: ‘Research Issues in the Study of Computer Supported Collaborative Writing’, in Mike Sharples (Ed.): *Computer Supported Collaborative Writing*, Computer Supported Cooperative Work, pp. 9–28, Springer London, 1993. (Cited on page [12](#))
- Fatimah Sidi, Payam Hassany Shariat Panah, Lilly Suriani Affendey, Marzanah A. Jabar, Hamidah Ibrahim, and Aida Mustapha: ‘Data quality: A survey of data quality dimensions’, in: *Proceedings of the 2012 International Conference on Information Retrieval & Knowledge Management*, pp. 300–304, Kuala Lumpur, Malaysia, 2012. (Cited on page [2](#))
- Edgar A. Smith and R. J. Senter: *Automated Readability Index*, AMRL-TR-66-220, Aerospace Medical Research Laboratories, 1967. (Cited on pages [52](#) and [93](#))

- Thamar Solorio, Ragib Hasan, and Mainul Mizan: ‘A Case Study of Sockpuppet Detection in Wikipedia’, in: *Workshop on Language Analysis in Social Media (LASM) at NAACL HLT 2013*, pp. 59–68, Association for Computational Linguistics, Atlanta, Georgia, 2013. (Cited on page 30)
- Vicki Spandel: *Creating Writers: 6 Traits, Process, Workshop, and Literature*, Pearson, Saddle River, NJ, USA, 6th edition, 2012. (Cited on pages 52 and 62)
- Besiki Stvilia, Les Gasser, Michael B. Twidale, and Linda C. Smith: ‘A Framework for Information Quality Assessment’, *Journal of the American Society for Information Science* 58 (12): 1720–1733, 2007. (Cited on pages 49, 58, 61 and 120)
- Besiki Stvilia, Michael B. Twidale, Linda C. Smith, and Les Gasser: ‘Information quality discussions in wikipedia’, *Technical report*, University of Illinois at Urbana Champaign, 2005. (Cited on page 119)
- Besiki Stvilia, Michael B. Twidale, Linda C. Smith, and Les Gasser: ‘Information quality work organization in wikipedia’, *Journal of the American Society for Information Science and Technology* 59 (6): 983–1001, April 2008. (Cited on pages 50, 58, 61, 110, 120 and 151)
- Karl-Erik Sveiby: ‘Transfer of knowledge and the information processing professions’, *European Management Journal* 14 (4): 379–388, 1996. (Cited on page 46)
- Zareen Syed and Tim Finin: ‘Unsupervised techniques for discovering ontology elements from Wikipedia article links’, in: *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pp. 78–86, Los Angeles, CA, USA, 2010. (Cited on page 25)
- David Tax: *One-class classification*, Phd dissertation, Technische Universiteit Delft, 2001, Online: <http://homepage.tudelft.nl/n9d04/thesis.pdf>. (Cited on pages 76 and 77)
- David R. Traum: ‘20 questions on dialogue act taxonomies’, *Journal of Semantics* 17: 7–30, 2000. (Cited on page 113)
- Grigorios Tsoumakas and Ioannis Katakis: ‘Multi-label classification: An overview’, *International Journal of Data Warehousing and Mining* 3 (3): 1–13, 2007. (Cited on page 146)
- Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas: ‘Mining Multi-label Data’, in Oded Maimon and Lior Rokach (Eds.): *Data Mining and Knowledge Discovery Handbook*, chapter 34, pp. 667–685, Springer, 2010. (Cited on page 167)
- Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave: ‘Studying Cooperation and Conflict Between Authors with History Flow Visualizations’, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 575–582, Vienna, Austria, 2004. (Cited on page 115)

- Fernanda B. Viégas, Martin Wattenberg, Jesse Kriss, and Frank Ham: ‘Talk Before You Type: Coordination in Wikipedia’, in: *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, pp. 78–78, Big Island, HI, USA, 2007. (Cited on pages [35](#), [114](#), [115](#) and [122](#))
- Richard Y. Wang and Diane M. Strong: ‘Beyond accuracy: What data quality means to data consumers’, *Journal of Management Information Systems* 12 (4): 5–33, March 1996. (Cited on pages [47](#), [48](#), [50](#), [59](#), [60](#), [63](#) and [156](#))
- Dennis M. Wilkinson and Bernardo A. Huberman: ‘Cooperation and Quality in the Wikipedia’, in: *Proceedings of the International Symposium on Wikis and Open Collaboration (WikiSym '07)*, pp. 157–164, Montreal, Canada, 2007. (Cited on page [66](#))
- Michael J Wise: ‘YAP3: Improved Detection Of Similarities In Computer Program And Other Texts’, in: *Proceedings of the twenty-seventh SIGCSE technical symposium on Computer science education*, pp. 130–134, Philadelphia, PA, USA, 1996. (Cited on page [105](#))
- Ian Witten, Eibe Frank, and Mark Hall: *Data mining : Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Burlington, MA, 2011. (Cited on page [94](#))
- Eti Yaari, Shifra Baruchson-Arbib, and Judith Bar-Ilan: ‘Information quality assessment of community generated content: A user study of Wikipedia’, *Journal of Information Science* 37 (5): 487–498, August 2011. (Cited on page [49](#))
- Yiming Yang and Jan O Pedersen: ‘A Comparative Study on Feature Selection in Text Categorization’, in: *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 412–420, San Francisco, CA, USA, 1997. (Cited on page [147](#))
- Torsten Zesch and Iryna Gurevych: ‘Analysis of the Wikipedia Category Graph for NLP Applications’, in: *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007)*, pp. 1–8, Rochester, NY, USA, 2007. (Cited on page [25](#))
- Torsten Zesch, Christof Müller, and Iryna Gurevych: ‘Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary’, in: *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco, 2008. (Cited on pages [40](#) and [163](#))

Index

- 1001 Nights Cast, [15](#)
- accessibility quality, [47](#)
- administrative pages, [23](#)
- algorithm adaptation, [146](#)
- Article Feedback Tool, [56](#)
- article naming conventions, [24](#)
- Assertives/Representatives, [112](#)
- bazaar model, [10](#)
- category graph, [25](#)
- category system, [25](#)
- cathedral model, [10](#)
- Claude Shannon, [45](#)
- cleanup templates, [28](#), [57](#), [67](#), [68](#)
- ClearTK, [89](#), [90](#)
- closed group collaboration, [9](#)
- coherence, [53](#)
- cohesion, [53](#)
- collaboration support system, [11](#)
- collaborative writing, [12](#)
- collaborative writing system, [15](#)
- collaborative writing tool, [15](#)
- Commissives, [112](#)
- communicative act, [113](#)
- computer supported cooperative work, [9](#)
- constructive hypertext, [15](#)
- constructive narratives, [14](#)
- content pages, [23](#)
- contextual quality, [47](#)
- conversation act, [113](#)
- conversational move, [113](#)
- cut-and-paste procedure, [36](#)
- Data Machine, [40](#)
- data-driven IQ management, [2](#)
- Declarations, [112](#)
- demarcation criteria, [21](#)
- dialog, [113](#)
- dialog act, [113](#), [132](#)
- dialog move, [113](#)
- Diff, [31](#)
- DiffAction, [126](#)
- DiffPage, [32](#)
- Directives, [112](#)
- disambiguation page, [24](#)
- discourse segmentation, [121](#)
- discourse structure, [121](#)
- discussion pages, [23](#)
- distinguished content certification, [54](#)
- DKPro TC, [87](#), [165](#)
- edit, [31](#)
- encyclopedic content, [23](#)
- entropy, [45](#)
- explicit threading, [34](#)
- exploratory hypertext, [14](#)
- Expressives, [112](#)
- featured article criteria, [54](#)
- featured content, [54](#)

five pillars, 21
flagged revisions, 33, 57
FlawFinder, 85, 165
founding principles, 21

generic information quality framework, 46
glossaries, 26
good articles, 54
group single-author writing, 13

Hadoop, 41, 81
hatnotes, 27
help pages, 23
horizontal division writing, 13
human-human dialog, 113
Hypertext Hotel, 15

illocutionary act, 112
in-text replies, 122
indexes, 26
infobox, 26
information, 45
information content, 45
information quality framework, 46
information quality model, 46
inline scope, 68
inter-page organization, 22
Interpedia, 20
interwiki links, 27
intra-page organization, 22
intrinsic article quality, 60
intrinsic quality, 47
IQ assessment, 50
IQ assurance, 50
IQ improvement, 50
IQ management, 50

Jimmy Wales, 20
John Austin, 112
JWPL, 40, 80, 89, 91, 120, 130, 132, 146

label reliability, 71
language version, 20
linguistic quality, 51
Lists, 26
locutionary act, 112
longest common substring, 162
Lua, 28

macrostructure, 22
Mallet, 90, 167
MapReduce, 82
Microsoft Office 365, 16
microstructure, 22
model of communication, 45
monologue, 113
motivation, 10
move procedure, 36

namespace, 23
Nupedia, 20

one-class classification, 77
online encyclopedia, 19
open collaboration, 10
open collaborative writing, 14
outlines, 26
overviews, 26

page blanking, 126
page scope, 68
PAN Challenge, 74
paragraph blanking, 126
parallel writing, 13
partial turns, 127
peer review, 55
pending changes, 33, 57
perlocutionary act, 112
portals, 26
power law of participation, 11
problem transformation, 146
process-driven IQ management, 2

PU learning, 77
 quality assessment, 65
 quality flaw detection, 70
 quality flaws, 68
 quality management, 12
 quality problems, 50, 67
 quality standard, 46
 reactive writing, 13
 readability, 52
 redirects, 25
 redlinks, 26, 27
 reliable negative, 81
 reliable positive, 82
 representational quality, 47
 requirements of content, 58
 requirements of demand, 58
 requirements of form, 58
 requirements of the project, 58
 revision, 31
 revision history, 31, 161
 revision of origin, 124
 RevisionMachine, 40, 161
 Scribunto, 28
 section scope, 68
 semi-structured resource, 22
 sequential writing, 13
 sock puppets, 30
 soft security, 30
 soft threading, 34
 standard-by-comparison, 54
 stratified division writing, 13
 SWEBLE, 89
 systemic bias, 22, 30
 Talk namespace, 23
 Talk pages, 33, 57
 template substitution, 28
 template transclusion, 28
 templates, 28, 36
 text quality, 51
 The Unknown, 15
 TimeMachine, 40, 161
 topic bias, 78
 topical preference, 70
 topical restriction, 70
 trade-off relations, 49
 turn, 34, 113
 turn-taking, 113
 UIMA, 87, 146, 166
 unstructured resource, 22
 user pages, 23
 user status groups, 29
 vandalism, 126
 Visual Editor, 28
 Weka, 90, 146, 167
 wiki, 16
 wiki markup, 27
 Wikidata, 27, 28, 42, 74
 WikiHadoop, 41, 81
 wikilinks, 26
 Wikimedia Labs, 39
 Wikimedia Toolserver, 39
 Wikipedia, 15, 16, 19, 53, 161
 Wikipedia Foundations, 21
 Wikipedia Miner, 40
 Wikipedia Protection Policies, 32
 WikiProject, 26, 31
 WikiProjects article quality grades, 66, 67
 Wiktionary, 41
 work coordination, 11, 33
 writing, 12
 writing quality, 51
 writing tool, 16
 writing traits, 51

Wissenschaftlicher Werdegang des Verfassers[¶]

| | |
|---------------|---|
| 2004–2009 | Studium für das Lehramt an Gymnasien Hauptfächer Informatik und Englisch Julius-Maximilians-Universität Würzburg |
| 2005–2009 | Magisterstudium Hauptfach Englische Sprachwissenschaft Nebenfächer Informatik und Englische Literaturwissenschaft Julius-Maximilians-Universität Würzburg |
| Juni 2009 | Abschluss als Magister Artium Magisterarbeit: „Semantic Relations in WordNet and the BNC“ im Bereich Englische Sprachwissenschaft Referent: Prof. Dr. Ilka Mindt |
| Dezember 2009 | Abschluss des 1. Staatsexamens für das Lehramt an Gymnasien |
| 2010–2013 | Wissenschaftlicher Mitarbeiter am Ubiquitous Knowledge Processing Lab Technische Universität Darmstadt |
| seit 2014 | Wissenschaftlicher Mitarbeiter am Ubiquitous Knowledge Processing Lab Deutsches Institut für Internationale Pädagogische Forschung, Frankfurt am Main |

Ehrenwörtliche Erklärung[‡]

Hiermit erkläre ich, die vorgelegte Arbeit zur Erlangung des akademischen Grades „Doktor-Ingenieur (Dr.-Ing.)“ mit dem Titel „*The Quality of Content in Open Online Collaboration Platforms: Approaches to NLP-supported Information Quality Management in Wikipedia*“ selbstständig und ausschließlich unter Verwendung der angegebenen Hilfsmittel erstellt zu haben. Ich habe bisher noch keinen Promotionsversuch unternommen.

Darmstadt, den 15. Mai 2014

Oliver Ferschke, M.A.

[¶] Gemäß § 20 Abs. 3 der Promotionsordnung der Technischen Universität Darmstadt.

[‡] Gemäß § 9 Abs. 1 der Promotionsordnung der Technischen Universität Darmstadt.

Publikationsverzeichnis des Verfassers

Johannes Daxenberger, **Oliver Ferschke**, Iryna Gurevych and Torsten Zesch: ‘DKPro TC: A Java-based Framework for Supervised Learning Experiments on Textual Data’, in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, pp. 61–66, Baltimore, MD, USA, June 2014.

Lucie Flekova, **Oliver Ferschke** and Iryna Gurevych: ‘What Makes a Good Biography? Multidimensional Quality Analysis Based on Wikipedia Article Feedback Data’, in: *Proceedings of the 23rd International World Wide Web Conference (WWW 2014)*, pp. 855–866, Seoul, Korea, April 2014.

Oliver Ferschke, Iryna Gurevych and Marc Rittberger: ‘The Impact of Topic Bias on Quality Flaw Prediction in Wikipedia’, in: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 721–730, Sofia, Bulgaria, August 2013.

Oliver Ferschke, Johannes Daxenberger and Iryna Gurevych: ‘A Survey of NLP Methods and Resources for Analyzing the Collaborative Writing Process in Wikipedia’, in Iryna Gurevych and Jungi Kim (Eds.): *The People’s Web Meets NLP: Collaboratively Constructed Language Resources*, Chapter 5, pp. 121–160, Springer, April 2013.

Oliver Ferschke, Iryna Gurevych and Marc Rittberger: ‘FlawFinder: A Modular System for Predicting Quality Flaws in Wikipedia - Notebook for PAN at CLEF 2012’, in: *CLEF 2012 Labs and Workshop, Notebook Papers*, Online Proceedings, Rome, Italy, September 2012.

Johannes Daxenberger, **Oliver Ferschke** and Iryna Gurevych: ‘Wikipedia-based Corpora for Analyzing Revisions, Discussions and Text Quality in Collaborative Writing’, in: *Workshop on Automatic Processing of Non-Standard Data Sources in Corpus-Based Research* (Extended Abstract), Cologne, August 2012.

Oliver Ferschke, Iryna Gurevych und Yevgen Chebotar: ‘Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages’, in: *Proceedings of the 13th Conference of the European Chapter of the ACL (EACL 2012)*, pp. 777–786, Avignon, France, April 2012.

Oliver Ferschke, Torsten Zesch and Iryna Gurevych: ‘Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia’s Edit History’, in: *Proceedings of the 49th Annual*

Meeting of the Association for Computational Linguistics: Human Language Technologies. System Demonstrations, pp. 97–102, Portland, OR, USA, June 2011.

Oliver Ferschke. ‘Semantic Relations in WordNet and the BNC’, M.A. Thesis, Universität Würzburg, Januar 2009.

