2013

Predicting Text Quality: Metrics for Content, Organization and Reader Interest

Annie Louis
University of Pennsylvania, annieplouis@gmail.com

Predicting Text Quality: Metrics for Content, Organization and Reader Interest

Abstract
When people read articles---news, fiction or technical---most of the time if not always, they form perceptions about its quality. Some articles are well-written and others are poorly written. This thesis explores if such judgements can be automated so that they can be incorporated into applications such as information retrieval and automatic summarization.

Text quality does not involve a single aspect but is a combination of numerous and diverse criteria including spelling, grammar, organization, informative nature, creative and beautiful language use, and page layout. In the education domain, comprehensive lists of such properties are outlined in the rubrics used for assessing writing. But computational methods for text quality have addressed only a handful of these aspects, mainly related to spelling, grammar and organization. In addition, some text quality aspects could be more relevant for one genre versus another. But previous work have placed little focus on specialized metrics based on the genre of texts.

This thesis proposes new insights and techniques to address the above issues. We introduce metrics that score varied dimensions of quality such as content, organization and reader interest. For content, we present two measures: specificity and verbosity level. Specificity measures the amount of detail present in a text while verbosity captures which details are essential to include. We measure organization quality by quantifying the regularity of the intentional structure in the article and also using the specificity levels of adjacent sentences in the text. Our reader interest metrics aim to identify engaging and interesting articles. The development of these measures is backed by the use of articles from three different genres: academic writing, science journalism and automatically generated summaries. Proper presentation of content is critical during summarization because summaries have a word limit. Our specificity and verbosity metrics are developed with this genre as the focus. The argumentation structure of academic writing lends support to the idea of using intentional structure to model organization quality. Science journalism articles convey research findings in an engaging manner and are ideally suited for the development and evaluation of measures related to reader interest.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Computer and Information Science

First Advisor
Ani Nenkova

Keywords
readability, text quality, writing quality

Subject Categories
Computer Sciences

PREDICTING TEXT QUALITY: METRICS FOR CONTENT, ORGANIZATION AND READER INTEREST

Annie Priyadarshini Louis

A DISSERTATION

in

Computer and Information Science

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2013

Supervisor of Dissertation

Signature _____

Ani Nenkova, Assistant Professor, Computer and Information Science

Graduate Group Chairperson

Signature _____

Val Tannen, Professor, Computer and Information Science

Dissertation Committee

Aravind Joshi, Professor, Computer and Information Science (Chair)

Mitch Marcus, Professor, Computer and Information Science

Ben Taskar, Associate Professor, Computer and Information Science

Regina Barzilay, Associate Professor, Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology (External)

Hal Daumé III, Assistant Professor, Department of Computer Science, University of Maryland (External)

PREDICTING TEXT QUALITY: METRICS FOR CONTENT, ORGANIZATION AND READER INTEREST

To

my mom, dad and sister

Acknowledgements

Numerous people have helped me on my way to a successful education. I would like to thank and acknowledge all of them.

First, I want to thank my advisor Ani Nenkova. She patiently taught me how to choose good research problems, to turn my ideas into measurable results, and to think critically and improve my work. Apart from research, she guided me through different aspects of scholarly life—giving talks, reading and writing papers, and developing my own opinions about research directions. I am thankful for her faith in my abilities and for encouraging me to never give up. Despite her busy schedule, she shared her knowledge and time with me generously for the last five years. I am going to miss our weekly meetings.

I am also very grateful to Aravind Joshi for his collaboration and kind mentorship. His wisdom, enthusiasm for research, and advice shaped my graduate student life in very favorable ways. I thank you for the support you provided me from my first days at Penn. I am also lucky to be surrounded by other wonderful faculty at Penn. I have enjoyed the many interesting conversations with Mitch Marcus, who is knowledgeable on practically any topic. Ben Taskar and Lyle Ungar provided comments and suggestions for my research thoughout my time at Penn.

I also thank Regina Barzilay and Hal Daumé III, my external committee members, for reading my thesis and providing valuable feedback.

Many others also had an influence on my research work. I acknowledge the support I received from Bonnie Webber during her visits to Penn. Bonnie spent several hours with me discussing my research, asking questions and suggesting future directions. I also thank Derrick Higgins from ETS who mentored me during my first internship in the industry and introduced me to research work outside academia. Thanks also to my collaborators during my internships at Yahoo Labs!—Jean François Crespo and Eric Crestan, and Microsoft Research—Todd Newman and Lili Cheng.

I am also grateful for the friendship and collegiality of the Penn NLP and machine learning groups. I learned so much from all my fellow students. Emily Pitler and I

started graduate school at the same time and Emily has been a collaborator and friend to me. She has listened to almost every practice talk I gave during these years and was the 'go to' person when I panicked about shortening my 'always too long' drafts of papers. Qiuye Zhao is a dear friend who was so much fun to talk to and made me wish she would work in the office more often. I also thank Constantine Lignos, Houwei Cao, Jennifer Gillenwater, Junyi Li, Paramveer Dhillion, Thomas François and Xi Lin for their friendship and help. My first NLP/ML officemates at Penn—Axel Bernal, Nikhil Dinesh and Ted Sandler—deserve special thanks for their encouragement during my early years.

I also appreciate the help I have received on numerous occasions from the administrative assistants of the Computer Science department and the Moore Business Office. Special thanks to Mike Felker, one of the most efficient people I know.

At this point in my educational career, I would also like to thank my professors and teachers from my earlier school days who equiped me with the necessary skills to take on further intellectual pursuits. I thank all my professors at my undergraduate school—College of Engineering, Chennai. I also thank my middle and high school teachers at St. Joseph's, St. Francis, and Stanes high schools in India. I fondly appreciate Mrs. Asuntha who made me more confident of my abilities, Mrs. Mary Van Haltren who brought me out of my shell and encouraged me to participate in many competitions, and Mrs. Mary Immaculate who pushed me to always aim higher.

At a personal level, my many friends made my hectic graduate student life more enjoyable and fun. I thank Sudeepa Roy and Marie Jacob, my closest pals in graduate school. I will miss our coffee breaks and numerous late dinners in the department. I wish you all success in your careers. I also thank Nan Zheng, Medha Atre, Marzie Taheri Sanjani and Yu-Wen Liu for their companionship. My friends from far away, Neetu Chajed and Malathi Raghavan, also deserve my heartfelt gratitude for their continuing support. I have also enjoyed the time and conversations with other friends at Penn—Brent Yorgey, Elena Bernardis, Julia Stoyanovich, Ling Ding, Mengmeng Liu, Michael Greenberg, Mukund Raghothaman, Partha Talukdar, Sanjian Chen, Santhosh Nagarkatte, Varun Aggarwala, Vincent Wang, and Zhuowei Bao.

I would not have been able to achieve any of these successes without my family—

my dad, mom and sister. My mom is my closest confidant and friend. She never had the slightest doubt in my abilities and always believed I could achieve anything. Her encouragement, love and optimism kept me moving forward in difficult times. My dad kept me in the company of many students and intellectuals and encouraged me to be ambitious from my childhood. I thank my parents for instilling in me a love for learning new things, and for teaching me the values of hard work and modesty. They have made numerous sacrifices to give me a good education and supported every decision I have made in my professional life. My dear sister, Annie (yes, we share the same name) has taken care of me so much and thinks about my needs even before I realize them. Annie, one of the wonderful things about finishing this thesis is that I am moving closer to where you live and we will be able to spend more time together, in person rather than on Skype. The sacrifices, prayers and love of my family have been the backbone for my work and success. This thesis is dedicated to you.

ABSTRACT

PREDICTING TEXT QUALITY: METRICS FOR CONTENT, ORGANIZATION AND READER INTEREST

Annie Priyadarshini Louis

Ani Nenkova

When people read articles—news, fiction or technical—most of the time if not always, they form perceptions about its quality. Some articles are well-written and others are poorly written. This thesis explores if such judgements can be automated so that they can be incorporated into applications such as information retrieval and automatic summarization.

Text quality does not involve a single aspect but is a combination of numerous and diverse criteria including spelling, grammar, organization, informative nature, creative and beautiful language use, and page layout. In the education domain, comprehensive lists of such properties are outlined in the rubrics used for assessing writing. But computational methods for text quality have addressed only a handful of these aspects, mainly related to spelling, grammar and organization. In addition, some text quality aspects could be more relevant for one genre versus another. But previous work have placed little focus on specialized metrics based on the genre of texts.

This thesis proposes new insights and techniques to address the above issues. We introduce metrics that score varied dimensions of quality such as content, organization and reader interest. For content, we present two measures: specificity and verbosity level. Specificity measures the amount of detail present in a text while verbosity captures which details are essential to include. We measure organization quality by quantifying the regularity of the intentional structure in the article and also using the specificity levels of adjacent sentences in the text. Our reader interest metrics aim to identify engaging and interesting articles. The development of these measures is backed by the use of arti-

cles from three different genres: academic writing, science journalism and automatically generated summaries. Proper presentation of content is critical during summarization because summaries have a word limit. Our specificity and verbosity metrics are developed with this genre as the focus. The argumentation structure of academic writing lends support to the idea of using intentional structure to model organization quality. Science journalism articles convey research findings in an engaging manner and are ideally suited for the development and evaluation of measures related to reader interest.

Contents

List of Tables

List of Illustrations

Chapter 1

Introduction

On a regular basis, we encounter some poorly-written articles and other well-written ones. This perception of quality is influenced by numerous factors: interesting topic, informative content, no errors in grammar and spelling, clear organization, elegant writing and also good layout and presentation of text on the page. As humans, we are able to make such judgements spontaneously. This thesis explores how we can automate judgements of text quality.

There are several situations where we would like to automatically measure the quality of an article.

Writing assessment is one area which can benefit from such systems. Teachers have to regularly grade student essays and provide writing feedback. In addition to classroom settings, large scale testing such as Graduate Record Examination[1], Test of English as a Foreign Language[2] and SAT[3] also involve rating thousands of student essays. For such situations, an automatic method of assessment can provide ratings that have greater consistency compared to those that can be assigned by people. Automatic assessment is also much cheaper than manual grading. This motivation has led to the development of commercial systems which can detect issues with spelling, grammar, and organization elements of student essays. For example, Educational Testing Service[4], a leading provider

[1] http://www.ets.org/gre
[2] http://www.ets.org/toefl
[3] http://sat.collegeboard.org/home
[4] http://www.ets.org

of high-stakes tests, has introduced an automatic essay grading system called *e-rater* [4].

Text quality prediction is also useful for recommendation and retrieval of articles. Today, practically any query issued to web search engines returns thousands of relevant results. However, although a huge number of articles are present on the web, not all articles are well-written. So in addition to query relevance, it would be helpful for users if the top results are also articles that are well-written.

Another big area of impact is the development of automatic summarization and generation systems. Even for the most mature genre of news summarization, systems have become good at selecting important content but the linguistic quality of generated summaries is rather poor. Text quality measures can help these systems score their hypotheses and create coherent and well-formed text. In addition, text quality prediction is useful for system evaluation. Consider for example, the summarization evaluations organized by NIST every year called the Text Analysis Conference (TAC[5]). There are about 50 systems that participate every year and each system produces summaries for about 50 test articles. These summaries are manually rated by assessors for content and linguistic quality and involves huge time and cost investment. While such evaluations are possible in focused workshops, researchers find it difficult to evaluate systems during development. Automatic metrics for linguistic quality will fill this gap.

Given these motivations, this thesis supplies a framework and a suite of automatic metrics for text quality prediction. The first challenge for this task is how to define text quality. Quality is a coarse concept and can mean different things to people—free of errors, interesting, informative, and well-organized, to name a few. Rather than picking one of the definitions, we adopt criteria proposed by education specialists in the form of rubrics for grading writing. These rubrics are widely accepted in the education community and are standardly used by school teachers as guidelines for writing assessment. Based on these criteria, text quality can be viewed as comprising four coarse concepts–conventions, ideas/content, organization and reader interest. Conventions are related to the mechanics of a language such as spelling, punctuation and grammar. Ideas refer to the choice of topic and subject matter that is presented. Organization gives structure

[5]http://www.nist.gov/tac/

to the content and presents it in an optimal sequence. A writer also adds his personal touch to the article to make it engaging, vivid and interesting. These factors constitute the reader interest concept.

When we analyze prior computational work according to these rubrics, most studies fall into only two of the dimensions—conventions and organization. The reason for this skew is differing interests of research groups that work on the topic. From the education side, metrics focused on spelling, preposition and article errors are most useful for assessment of writing because learners and non-native speakers of a language are the large target groups. Similarly, a large number of organization-related metrics have been proposed with automatic summarization as a target application. Particularly, in multi-document summarization, content is chosen from multiple source documents and needs to be properly ordered in the output summary. This need has motivated the development of methods that learn which topic, entity and discourse relation transitions can distinguish well-organized texts from other incoherent examples. However, in several other application settings, these metrics are inadequate. For example, recent years have seen the release of large archives of news articles such as New York Times[6] and Google News[7] archives. These resources provide excellent opportunities for search and browsing applications. But since these articles are written by professional journalists, conventions-based metrics are of little use for predicting their quality.

The other issue that is unexplored so far is the role of genre. Good writing involves different criteria depending on its genre. For example, fast-paced storytelling makes fiction interesting whereas clarity and explanations make technical articles well-written. Therefore using texts from diverse genres, we can develop text quality metrics that capture many different quality dimensions. On the other hand, we would also like to have metrics that are predictive and have stable accuracies across multiple genres. But little focus has been given in previous work for genre-based analyses. Previous metrics were developed based on only a few genres and the metrics were mostly evaluated only on texts from one genre. Part of the problem is the non-availability of suitable corpora for many genres. Essays written by non-native speakers and news summaries created by

[6]http://www.nytimes.com/ref/membercenter/nytarchive.html/

[7]news.google.com/archivesearch/

automatic systems are the large scale datasets available with ratings for writing quality. So metrics for conventions are geared towards errors made by non-native speakers and organization-metrics towards informational texts. But today we see that for tasks such as spelling correction, techniques that utilize domain-knowledge are better and sometimes necessary, for example, in spelling correction of search queries or email. Similarly, the needs of summarization systems are expanding, and fiction, conversations, and academic writing genres are attracting interest within the summarization community. Ideally, text quality metrics should involve generic aspects that are stable across genres and domain-specific measures predictive of quality for individual types of text. But this breadth is little achieved by current work.

This thesis is a step towards addressing the two issues above and presents a better prediction model for text quality.

We introduce new metrics for the unexplored aspects of content and reader interest and also introduce metrics for organization quality. Our metrics measure the specificity of content, verbosity level, "interesting" nature of writing and regularity of intentional structure. The idea of specificity is based on a two level distinction: general topic-related information and specific details. Good writing has a proper balance between general and specific content. Our metric associates a specificity level with each sentence of an article, enabling us to study how the level of specificity in the article impacts text quality. Our verbosity measure is also related to the details presented but instead determines whether the details are appropriate and at the right level for the articles. While discussed much in style manuals on writing, there have been no attempts to automatically predict verbose writing. We propose a data-driven approach for verbosity prediction which relies on learning a mapping between the type of content included and length of articles. These two metrics capture content quality. From the reader interest perspective, we introduce metrics to identify articles that are perceived as engaging to readers. These metrics are based on word choice, creative language use and structure of the article. Finally, we also present a measure for organization quality based on the intentional structure of an article. This method is based on the idea that each sentence in an article can be associated with a communicative goal from the point of view of the author. Some sequences of commu-

nicative goals work better to convey the overall message compared to others. Our metric aims to capture this aspect by assigning scores to texts based on the sequence of communicative goals present in them. The specificity metric also has ties to organization: we use the scores to capture the arrangement of general and specific information in articles.

In addition, this thesis presents the first work that investigates text quality prediction on different genres of text. We consider three genres: academic writing, science journalism and automatically produced summaries. This genre-based approach allows us to develop and test the wider range of metrics proposed in this thesis. We use articles from the science journalism genre to study interest-related measures. These articles convey research content to lay readers in an understandable and also entertaining manner. They even involve a storyline, humour and suspense elements, making them suitable for exploring this dimension. Similarly, the academic writing genre motivated the development of the intentional structure metric. Academic writing is commonly seen as having an argumentative structure where the researchers highlight why their proposed solution is important and useful. Our prediction method is unsupervised and does not require manual annotation of intentions for training. However, existing annotations of intentions in academic writing help us to motivate and test our approach. Finally, we consider summaries generated by automatic systems also as a genre on its own. Systems make very different errors than people and so quality metrics need to be modified appropriately. For example, an automatic summary may introduce a pronoun in a sentence without a proper antecedent in the earlier context. In addition, summaries have a length constraint, so information should be organized in a judicious manner to convey the important content. As a result, aspects such as how specific the content is and how it is presented should be useful for predicting the quality of automatic summaries and inspired our specificity and verbosity metrics.

1.1 Thesis organization

Chapter 2 describes how we define text quality in terms of rubrics used for writing assessment. We propose the Six Traits rubric from the education genre as a well-founded framework for text quality prediction. A particularly appealing character-

istic of this rubric is the diverse aspects covered by it, ranging from information quality and writing style to handwriting and layout of the page. We provide a review of prior work based on this definition of text quality and show how several of these traits have not been explored for computational work and how our metrics address this gap. We also discuss the issue of test data for evaluation. We describe the corpora and quality ratings that we use for the academic and summarization genres. For summarization, we utilize existing ratings from large scale evaluation workshops. For academic writing, we rely on approximate examples by manipulating articles to obtain incoherent examples.

Chapter 3 For science journalism, we introduce a new corpus of articles categorized for text quality. We collected samples of great writing from a popular anthology that publishes articles rated by expert journalists as outstanding and engaging science writing. We expanded this set of well-written articles with more samples written by authors whose articles appeared in the anthology. We created an opposite category of typical writing by collecting articles on similar topics as the great writing but not appearing in the anthologies or written by the authors of the great writing samples. We create two corpora from this data. One contains the categories above and the other groups articles by topic so that we can explore how to perform text quality prediction for articles with similar topic. This chapter provides the details about the corpus and also presents an annotation study comparing our corpus categories created on the basis of expert judgements with ratings provided by an adult reader who is not a journalist.

Chapter 4 introduces our metric for organization that relies on the intentional structure of articles. It uses the idea that every sentence in an article has a purpose associated with it and the sequence of sentences helps the author achieve his overall goal for the discourse. Our approach aims to assign an intention label to each sentence and examine the sequence of labels to compute a score depending on whether the sequence is a regular one for well-written articles. To naively implement this approach, we need to predefine the set of intentions (which would vary each time we change the genre of articles) and also create annotated data to build a super-

vised classifier for intentions. Our system overcomes this challenge by introducing an unsupervised approach for assigning intention-like labels. We use the syntax of the sentence as a rough proxy for its intention. Sentences from the training data are clustered by syntactic similarity to uncover categories and these approximate categories are used for the second part of examining sequences of intentions. We propose this metric as suitable for the genre of writing about research since such writing is often seen as an argument from the author. Our evaluations consist of a coherence prediction task for academic articles and a task to identify the text quality category for articles from our science journalism corpus. We show that both evaluations confirm the predictive strengths of this metric.

Chapter 5 presents the predictor that we developed for content specificity. We make the distinction between general content (overall ideas) and specific information. We build a sentence-level binary classifier and show that it has high accuracy for separating out sentences annotated by people for this distinction. We also show how the specificity scores provided by our classifier can be used to perform both content and linguistic quality evaluations for automatic summaries. In summarization, content quality so far is only evaluated by comparison to information present in human-written summaries and organization is evaluated on the basis of ordering of sentences. We show that other aspects such as the amount of details present in the summary are also significant indicators of both these dimensions of summary quality. Similarly, in science journalism, we hypothesize that the level of detail used for describing a research problem should influence quality. We use features related to text specificity to identify the categories in our science news corpus and find that they give performance above the baseline.

Chapter 6 presents a system specifically designed to do text quality prediction for science journalism. This study is the first to explore factors whch make an article interesting and engaging to readers. We present implementations of metrics related to visual nature, story-telling format, beautiful and surprising language use and amount of research descriptions and study how these measures are related to and indicative of the quality categories on the corpus that we have developed. We find that these

genre-specific measures provide high accuracy in text quality prediction. We also present comparison and combination of our features with measures proposed for assessing other writing aspects such as readability and well-written nature. We show that all these aspects are complementary and necessary for accurately distinguishing articles with different quality.

Chapter 7 presents an approach to measure verbosity of an article. We hypothesize that a binary decision of 'verbose or not' can be rather hard to make for an isolated text. So we design a measure with a graded scale. We propose an approximate model where we try to capture in a corpus of concise articles, the relationship between content type and article length. Using the model's behaviour on test articles, we identify whether an article has the appropriate level and type of details. We perform two evaluations of this measure—on automatic summaries and on science journalism articles. We show that assessment of content and linguistic quality for automatic summaries can be performed in a reliable manner using this measure. However, for science journalism, using features from our verbosity model did not provide much improvements above the baseline.

Chapter 8 summarizes the main ideas and contributions of this work. We also list some limitations of our methods and how they can be improved in future studies.

1.2 Thesis contributions

We make four main contributions:

Text quality based on linguistic indicators: We emphasize linguistic properties for text quality prediction which is a departure from previous work where the quality of being well-written is often conflated with "easy to read". These prior studies have sought to retrieve for a reader articles that would be comprehensible by him. We argue for lesser focus on the ability of an audience to understand a text and in contrast consider the task of identifying problems with writing through the lens of an expert reader. We provide a suitable definition of text quality based on standard and widely accepted rubrics for writing assessment by teachers which we show has

two desirable features—separation of writing quality from audience abilities and a framework covering many different dimensions of quality.

New approaches for predicting organization quality: We introduce two metrics for organization that are based on new insights about quality. We provide a method to score the specificity of a sentence and use it to explore how sequences of general and specific content contribute to proper organization of a text. We show that specificity is indicative of linguistic quality of automatic summaries and of the text quality categories on our science journalism corpus. A second metric captures how sentences with different intentions can be interleaved to convey the overall discourse purpose to a reader. This metric is one of the first designed based on the specific argumentative structure of research writing and our evaluations find it to be indicative of quality for both academic writing and science journalism articles.

New predictors for content and reader interest: Our work is the first to propose predictors for content quality and interesting nature of writing. We develop a classifier for predicting engaging articles versus average ones for the science journalism genre. We also propose specificity and verbosity measures which track the presentation of content in the article. Our results show that specificity of content is a significant predictor of content quality for automatic summaries. Our verbosity metric is also designed with the summarization task as a target and proves useful for both content and linguistic quality evaluation of summaries.

Genre-based study of metrics: We use articles from three different genres for our study. Firstly, this genre-based approach allows us to develop a wider range of metrics than previous work. Our content quality metrics are evaluated on summarization data and we use academic articles for evaluating the organization metric based on intentional structure. Our science journalism corpus provides the test data for metrics related to reader interest. Secondly, we are also able to test the stability of our metrics across genres. We find that content specificity as well as intentional structure regularity are significant predictors of quality for more than one genre. In addition, we introduce a new corpus of text quality tagged articles for the genre of

science journalism. A distinguishing feature of our corpus is that the articles are of high quality in general but are separated into categories based on the distinction of outstanding versus average. This corpus is one of the first to provide such finer level distinctions.

Chapter 2

Task specifics

This chapter presents the preliminaries for our work.

We start with the definition of text quality that we follow in this thesis (Section 2.1). Expectedly, people differ in their opinions about what constitutes good writing. So a definition of text quality should include only aspects that most people identify and use for their judgements. In addition, a person's background knowledge and preferences influence what they consider as well-written text. This issue requires us to also specify the target audience for whom the quality notion is relevant. In this thesis, we propose that scoring rubrics used by writing experts can address these problems in a reasonable and clear manner.

Particularly, we adopt the 'Six Traits' scoring rubric [150] as our definition of text quality. The Six Traits approach is widely used in the education sector to guide teachers for scoring student writing. It is based on empirical studies of how expert and adult readers grade student essays and includes the six aspects/traits of writing which most raters pointed out as critical and agreed upon. Today the Six Traits model is almost standard for writing assessment [150]. We use this rubric to set up the task and goal of text quality prediction: to automatically score writing based on these traits. As we will show, this definition provides a number of advantages for our task and also clearly separates our work from the related area of readability prediction.

In Section 2.2, we describe the main differences between our study and prior work. Particularly, we focus on readability studies where the competency of the audience plays

a pivotal role. Readability aims to select appropriate text for people with different age and education levels. In contrast, our goal is to have minimal impact of audience on our predictions. We do so by assuming an *expert* audience, at the highest level of competency. We argue that this choice enables us to focus on linguistic aspects of the text without conflating well-written nature with understandability of the text. For example, under readability, a newspaper article is appropriate for an adult reader but has low readability for a fifth grade student. But in our work, we assume that our audiences come from only one category, say college educated adult for all our texts, and further we consider that they are exposed only to articles appropriate for their reading level. In this setting, we want to identify which aspects of the text make these readers enjoy or dislike them. We also review prior work on predicting aspects of text quality and how our work differs from them. Most of these studies have focused on conventions and organization quality.

Section 2.3 describes the three genres used in this thesis. We discuss their special characteristics and quality issues that are likely to be present in each of them.

Another issue of utmost importance is the gold-standard test data for evaluating our metrics. Numerous difficulties are involved when creating corpora with text quality ratings. The cost for obtaining annotations is high and also annotators need to be suitably trained to identify the target quality aspect. Luckily, for one of our genres, automatic summaries, we have large scale datasets with manual ratings from annual summarization evaluation workshops. For academic writing, we create samples of low quality by manipulating well-written articles. We describe these details in Section 2.4. For science journalism, we present a new corpus that we have created which contains New York Times (NYT) articles divided into categories for text quality. We built this corpus in a semi-automatic manner by first using articles rated by expert journalists as good writing and expanding the corpus by adding other good and typical articles. Details about this corpus are given in the next chapter. For each corpus, we focus on two desirable features: relevance to our definition of text quality and audience level, and appropriateness for the individual aspect we wish to study using the corpus.

2.1 Defining text quality

To specify which aspects make up text quality, we propose to employ rubrics used to teach writing. Specifically, we focus on the Six Traits model [150] which enumerates six essential criteria for writing assessment.

The development of the Six Traits rubric was influenced by an early study by Diederich (1974) [34]. He used around fifty people in various capacities such as writers, editors, business executives and teachers and asked them to group 300 student essays into good, mediocre and poor quality categories. Specificially, the raters placed each essay in one of nine rating levels and wrote comments on the problems with each essay. There were two main findings from this study.

- People varied greatly about which class they assigned for each individual essay. However, there were groups of people who agreed with raters within the same group but disagreed with other groups. The study identified five groups and found that raters were often clustered according to their distinction as a teacher, business executive or professional writer.

- When the comments reported by people within each group were analyzed, it was found that each group focused greatly on a different aspect of quality. The main aspect considered as important by each group emerged as a basic set of criteria for good writing.

Later studies [110, 129] were also able to replicate these findings and all studies came up with a small (around five or six) definable set of traits for judging quality. These findings led a team of teachers to develop a rubric for grading writing covering the traits highlighted in prior work. This rubric is called the Six Traits [150] and is immensely popular and standard in the education field.

The traits (as excerpts from Spandel (2004) [150]) are shown in Tables 2.1 and 2.2. We also outline when a text is considered worthy of the highest score for each trait.

1. Ideas and development: The writing is clear, focused, and well-developed, with many important, intriguing details.

- The writer is selective, avoiding trivia, and choosing details that keep readers reading.
- Details work together to clarify and expand the main.
- The writer's knowledge, experience, insight or perspective lend the piece authenticity.
- The amount of detail is just right–not skimpy, not overwhelming.

2. Organization: The order, presentation, and structure of the piece are compelling and guide the reader purposefully through the text.

- The entire piece has a strong sense of direction and balance. Key ideas stand out.
- The structure effectively showcases ideas without dominating them.
- An inviting lead pulls the reader in, a satisfying conclusion provides a sense of closure.
- Details fit just where they are placed.
- Transitions are smooth, helpful and natural.
- Pacing is effective; the writer knows when to linger and when to move along.

3. Voice: The writer's passion for the topic drives the writing, making the text lively, expressive and engaging.

- The tone and flavor of the piece are well-suited to topic, purpose, and audience.
- The writing bears the clear imprint of this writer.
- The writer seems to know the audience and to care about their interests and informational needs.
- Narrative text is moving and honest; informational text is lively and engaging.
- This is a piece readers want to share aloud.

Table 2.1: Six Traits definition and criteria for very good essay - Part I (continued in Table 2.2)

4. Word choice: Precise, vivid, natural language enhances the message and paints a clear picture in the reader's mind.

- The writer's meaning is clear thoughout the piece.
- Phrasing is original–even memorable–yet the language is never overdone.
- Lively verbs lend the writing energy and power.
- Modifiers are effective and not overworked. Clichés, tired words, and jargon are avoided.
- The writer repeats words only for effect and does not overdo it.
- Striking words or phrases linger in the reader's memory

5. Sentence Fluency: Easy flow and sentence sense make text a delight to read aloud.

- Sentences are well-crafted, with a stong, varied structure that invites expressive oral reading.
- Striking variety in structure and length gives writing texture and interest.
- Purposeful sentence beginnings show how ideas connect.
- The writing has cadence as if the writer hears the beat in his/her head.
- Fragments, if used, add style and punch; dialogue, if used, is natural and effective.

6. Conventions: The writer shows excellent control over a wide range of age-appropriate conventions and uses them accurately–sometimes creatively–to enhance meaning

- Errors are so few and minor a reader could skip right over them unless searching for them.
- The text appears clean, edited, polished. It's easy to process.
- Only light touch-ups are needed before publication.
- Conventions enhance the message and voice.
- As appropriate, the writer uses layout to showcase the message.

Table 2.2: Six Traits definition and criteria for very good essay - Part II (continued from Table 2.1)

We define text quality prediction as the computational approach to detect and score these traits in writing. This definition allows us to clearly specify certain characteristics of the task.

1. **Range of traits:** The rubric highlights many different aspects that are considered essential and core for good writing. Even layout and presentation of the page plays a role in the conventions trait. For example, a webpage with good content and linguistic style can be obscured by bad formatting, font type and color choices. This wide focus of the rubrics provides motivation for text quality measures that cover different traits of writing to provide a better overall score for an article.

2. **Audience:** Another question for text quality prediction is "Who is the target audience whose quality perceptions we wish to model?" Audience can vary in age (child or adult), educational level (middle school student or college educated reader), technical expertise (expert researcher, novice in the field or lay public) and people with cognitive disabilities versus those without. The same text will be rated with different quality levels depending on the audience we choose. Even a well-written text may not be appreciated or understood by a reader with poor reading abilities. In this thesis, we assume an expert reader both in terms of content and reading proficiency. This setting gives us an exciting space to work with: maximum emphasis on the linguistic properties of the text without considering the abilities of a reader. The Six Traits rubric is an excellent example of our desired setting where teachers are experts and outlines all the deficiencies noticed by this expert reader. In many applications such as article recommendation and automatic summarization, the audience is often an expert. We are interested in obtaining text quality ratings with such a reader as the target.

For the purposes of this thesis, we will call 'voice', 'sentence fluency' and 'word choice' as one category—*reader interest*-related traits. These measures supplement already well-organized and error-free texts and make them engaging and interesting. Since this thesis is the first large scale evaluation and study of reader interest measures, we group these three aspects as one category but their individual characteristics are also worthy and important to explore in future.

2.2 Related work

We make our definition and task more specific by comparing them to prior work in the area. We also provide short descriptions of some of the techniques for quality prediction introduced in prior studies.

2.2.1 Readability

The largest area of work related to text quality is readability. Readability is defined as the ability of a reader to comprehend a given text. This concept has been widely studied from both psycholinguistic and computational perspectives.

On the computation side, the task of readability prediction is typically set up as follows:

Consider that there are audiences with different competency levels $R = (r_1, \ldots r_k)$. For example, we can think of the audiences as having different educational grade levels, say 1 to 12. We also have a pool of texts $T = (t_1, \ldots t_n)$ that are written for different levels. The task is to create a one-to-many mapping $R \rightarrow T$ which divides the pool of texts among the audience categories. In practice, some algorithms supply a probability distribution over the audience levels for each article rather than strict assignment to one of the levels [26].

There are different ways to define audience competency levels and mappings between texts and these levels. The most popular definitions are based on age, educational level and cognitive abilities.

One of the influential and early studies of readability was done by Flesch (1948) [49]. In this work the gold standard mappings are defined as follows: the readability score for a text is the average educational grade level of a child who after reading the text could answer 75% of comprehension questions based on the text's content. Flesch proposed two features which were significantly correlated with this readability score—average sentence length in words, and average word length in syllables. These two measures were combined to predict the reading score using the following formula (inverted to predict ease of reading rather than difficulty of text):

$$\text{Reading Ease} = 206.835 - .846wl - 1.015sl$$

Here *wl* indicates average word length and *sl* indicates average sentence length. Later work continued to use this setting of educational grade levels as the categories and reading material designed for these grades are taken as the test set. Other readability formula such as Gunning's Fog index [58] also use word and sentence lengths as the central components.

The familiarity of words is another factor computed and studied in several readability work. Unfamiliar words could contribute to reading difficulty. The Dale-Chall readability formula [30] is a popular example for the use of this idea and is based on a list of *familiar* words. The Dale-Chall list contains approximately 3000 words. It was constructed by examining several thousand words for whether they were familiar and understood by fourth grade students. They compute reading difficulty as:

$$\text{Reading difficulty} = 0.1579fa + .0496sl + 3.6365$$

where *fa* is the proportion of words in the text that are not present in the list of *familiar* words and *sl* is average sentence length. Later work has generalized such familarity lists through the use of language models. In these approaches [26, 147], a unigram language model is constructed on example texts from each grade level. The likelihood of a test article is computed using each of the models and the grade level corresponding to the model which gave the highest likelihood is taken as the predicted grade level.

Apart from words, syntactic complexity has also been shown to be indicative of reading difficulty. Schwarm and Ostendorf (2005) [144] incorporated scores related to sentence syntax together with traditional measures for sentence length, word length, familiarity and language model scores. They used machine learning to produce a prediction that combined these evidences. The four syntax features in their model were average parse tree height, the average number of noun phrases, the average number of verb phrases and average number of subordinate clauses.

The relationship between discourse properties of a text and reading ease has also been explored [7, 50, 124]. For example, Foltz, Kintsch and Landauer (1998) [50] studied the relationship between Latent Semantic Analysis (LSA) based sentence overlap scores and reader scores on comprehension tests. They found that greater cohesion and continuity

in a text was correlated with better recall of the text's subject matter by its readers. In this work, a large term document matrix was constructed and reduced to 300 dimensions. Each sentence in test article was represented by a vector which is composed from the reduced vectors for individual words. The similarity between the vectors of adjacent sentences is computed using cosine overlap and the average overlap value in the text is taken as its LSA overlap score.

In Feng, Elhadad and Huenerfauth (2009) [47], the task is to identify text that is appropriate for people with certain cognitive disabilities. They create a corpus of articles where for each article, they had their target users answer comprehension questions and the average score obtained by the users for each article was used as a measure of that article's difficulty. They follow a machine learning approach to predict these scores and incorporate specialized features to indicate differences that would be noticed by their target readership. Several of their features are based on number of entities and length of lexical chains since they are likely to be related to cognitive load while reading the article.

As reflected in the above studies, the central notion in readability is comprehension. For example, a fifth grade reader cannot understand a 12th grade text and hence the text is not appropriate for fifth grade level audience. In contrast, in text quality, we seek to remove the focus on the ability of the reader. We assume only texts from one reading level and we assume a fixed audience level. So we have readers and texts that they can comprehend well. The goal now is to see what these readers will perceive as well-written text. For example, most college-educated adult readers of a newspaper can understand the content presented in news articles. But they would consider some articles as more well-written and enjoyable compared to others. Similarly, a college student can understand a fifth grade text, it is 'readable' for him but would not necessarily be well-written. This departure from readability is a central feature of our work.

2.2.2 Metrics for conventions

The conventions trait defines properties for acceptable text in a language. Several studies have addressed the prediction of conventions quality focusing on spelling, grammar and punctuations.

Spelling and grammar correction tools are successful natural language processing (NLP) applications and have been commercially deployed in word processing software. There is also interest in addressing problems that arise in specific genres, for example, spelling and preposition errors made by non-native learners of English language [32, 53, 157], grammar errors specific to academic writing [31] and spelling correction for search queries [29, 83]. Educational testing enterprises have huge datasets of non-native writing which have been annotated for such errors and web search companies have query logs where spelling reformulations can be studied. These resources and commercial interest has led to a lot of work on this trait.

There is also work on layout and presentation of the text. Ivory and Hearst (2002) [66, 67] present studies on predicting the quality of webpages using features about their HTML layout. In that work, the ratings for web page quality are obtained from expert Internet professionals. Ivory and Hearst use features related to number of links, graphics, font type and size, text positioning and color to predict these ratings.

Since a wide range of conventions quality aspects have been explored and solved to some extent in prior work, in this thesis, we do not focus on conventions. In fact, we assume that the texts in our data sets have highest quality in this regard. We explain these assumptions further in Section 2.4.

2.2.3 Metrics for organization

Several theories have been put forth proposing surface cues in texts that help to tie sentences together and create a flow in the article. Halliday and Hasan (1976) [62] outline three properties—entity repetition, discourse relations and ellipsis. Grosz and Sidner (1986) [57] define intentional structure, entities and discourse segments as components of coherent organization. Centering theory [56] has focused on describing how entity repetition and pronoun use happen at the level of adjacent sentences of coherent articles.

Automatic metrics have been developed motivated by these theories. These metrics provide evidence that coreference and discourse relations can be used to predict coherence as hypothesized. Karamanis (2009) [71] computes the different entity transitions which are proposed by Centering Theory as necessary for coherent organization. Viola-

tions of these transitions are counted and used to assign a score for organization quality. They perform their experiments on two corpora where manually annotated coreference information was available [35, 127]. Pitler and Nenkova (2008) [124] use a language model of discourse relations to predict ratings of well-written nature. They perform their study on Wall Street Journal articles and use the discourse annotations available in the Penn Discourse Treebank [128] for building a unigram language model. For a test article, they compute the likelihood of discourse relations present in it as a multinomial probability:

$$P(T) = p(n)\frac{n!}{x_1!...x_k!}{p_1}^{x_1}...{p_k}^{x_k}$$

where text T has n total discourse relations. k is the number of relation types in the discourse framework. x_i indicates the number of times relation type i appears in the article and p_i is the probability of relation type i in the language model. $p(n)$ is the probability that an article contains n discourse relations. These approaches have fairly good accuracy, in fact, the sequence of discourse relations turn out to the most powerful class of predictors for ratings of well-written nature compared to other readability measures. However, these studies use oracle annotations of coreference and discourse relations since these annotations are still hard to automate with good accuracy.

In light of this problem, other methods have emerged which do not depend on such strict notions of linguistic structure. These approaches are data-driven and work with a corpus of naturally occurring texts from the chosen genre and require no further annotation. These methods learn word co-occurrences [8, 79, 149] and entity reference [7, 42, 45, 80] that is normal for adjacent sentences and use this knowledge to evaluate the flow of a new sequence of sentences during test time. For example, the Entity Grid method developed by Barzilay and Lapata (2008) [7] is also motivated by the Centering Theory idea of entity repetitions between adjacent sentences. However, in contrast to Karamanis work where Centering based transitions were explicitly computed, in the Entity Grid, patterns in entity transitions are learned from data.

Consider an example text containing three sentences $S1$ to $S3$.

(S1) The fairy appeared before the girl.

(S2) The girl wished to be freed from the giant.

	girl	fairy	wand	wish	giant
S1	X	S	—	—	—
S2	S	—	—	—	X
S3	—	S	O	O	—

Figure 2.1: An example Entity Grid representation

(S3) The fairy waved her wand and granted the wish.

To create an entity grid, a text is represented by a set of n rows corresponding to the sentences (here 3) and p columns one for each unique entity mentioned in the text. In our case, we would have five columns as shown in Figure 2.1.

The cell corresponding to the i^{th} sentence (row) and j^{th} entity (column) is filled with the corresponding entity's grammatical role (S-subject, O-object and X-other) in that sentence. The absence of entity j in sentence i is recorded by a '$-$' in cell ij. Figure 2.1 shows the populated Entity Grid for our example sentences. In this grid, a column's entries from top to bottom reflect that entity's transitions in the text. The entity *fairy* is the subject of the first sentence, absent in the second and reappears as the subject of the third. Therefore between any two adjacent sentences, different types of transitions SO, SX, $O-$, XX, $-X$, etc. can occur for the different entities. A total of $M = 16$ such transitions are possible including the $--$ transition.

Barzilay and Lapata record the total count of each of these transitions over the entire text. The proportion of each transition type among the total transitions is calculated. Each proportion is a feature in their model and the weights for these are learned in a discriminative setting for predicting articles with good and poor organization.

A similar approach was taken by Lin, Ng and Kan (2011) [90] who also use discourse information for predicting organization quality. Their work is based on the predictions of an automatic discourse relations parser compared to gold annotations used in Pitler and Nenkova's study. However, Lin, Ng and Kan fuse the discourse relation information into the entity grid framework and adopt data-driven learning of patterns rather than just use

the discourse relations.

This thesis proposes a new metric to score organization which takes the intentional structure of an article into account. Despite playing a central role in discourse theories, intentional structure has so far not been automated for predicting organization. We use an insight about the correlation between intention of a sentence and its syntactic structure and develop a model using a data-driven approach similar to those described above. Our genre-based approach has also been a motivation for modeling intentional structure. Academic writing in particular is mostly analyzed as an argument from the authors about their research [154, 160]. Therefore predicting organization quality based on intentional structure is ideally suited for this genre. We also introduce a metric that captures the degree of specificity of a sentence and allows us to check whether a sequence of sentences has preferred transitions between general and specific content.

2.2.4 Metrics for reader interest

As we move into reader interest and content measures, there is little prior work. Few efforts have been made to develop corpora or prediction methods for these aspects.

However, even one of the earliest readability work by Flesch [49] which we introduced in our review of readability studies (Section 2.2.1) proposes that reader interest is also important while computing readability scores. In that paper, Flesch studies four measures for each text. We discussed two of them in Section 2.2.1—average word length and average sentence length. He also computes two other measures: number of *personal* words (counts of pronouns referring to people and words like "people" and "folks"), and number of *personal* sentences (quotes, exclamations, questions, commands, requests and incomplete sentences whose meaning must be inferred from the context). The last two components are assumed to be related to "human interest" based on the idea that articles about people would be interesting to a reader. Flesch hypothesized that readers will be motivated to read more interesting articles subsequently leading to better comprehension. But since their gold standard was based on results from reading comprehension tests, he found that the human interest scores while correlated with the gold standard did not improve the correlation when combined with word and sentence length features. Later

work paid less attention to the "human interest" dimensions and today only the word and sentence length components are standardly used as the Flesch score.

But in recent work, McIntyre and Lapata (2009) [104] note that in practical situations interest measures become necessary. In this work, they create a knowledge base of entity and event co-occurrences by automatically learning these patterns from a corpus of fairy tales. They generate new stories using this information. Apart from choosing the entities and likely events associated with the entities, it is also important to maintain good linguistic quality for the stories. They compute a Entity Grid score for the organization of the story but also consider that interest measures should be included. Since their genre is stories, they suppose that the interest value of the generated story is also important. They obtain user ratings for interest on a small corpus of fairy tales and compute several token based scores related to part of speech tags, syntactic relations, and categories from the MRC psycholinguistic database [171]. They found that a supervised classifier based on these scores made accurate predictions of the user ratings. Number of objects, nouns and imagery related words turned out to be the features that had highest correlation with user ratings. When the stories were generated by optimizing for the interest metric, people liked the output stories more compared to stories which did not consider this aspect. This work is the first to our knowledge that models reader interest but in a preliminary manner.

In this thesis, we aim to study the voice, word choice and sentence fluency traits in science journalism, a genre suitable for analyzing this aspect of quality.

2.2.5 Metrics for content

The quality of content and interest value attached to topics is another relatively unexplored trait and is probably the most sophisticated of all. Several aspects of content quality such as the choice and importance of ideas are difficult to model using surface features.

But in certain task settings such as automatic summarization, content quality easier to define and appropriate gold standards have been developed. Summaries should contain the most important content from the input. Standard methods for evaluating the content

of summaries involves comparing the summary's content with that included by a human while producing a summary for the same source document. The comparison is done either manually using techniques such as content coverage scores[8] and pyramid evaluation [114] or automatically using ngram overlaps between the system and gold standard summary [86].

In other situations, approaches for predicting content quality rely on meta-content properties instead. For example, Burstein et al. (2003) [14] present a supervised approach to identify the main idea and supporting details in student essays since they are important for the argument of the essay. But they do not provide any text quality evaluation using them.

Summary evaluation techniques have not considered such meta-content aspects and measures related to how the content is presented in the summary. In our work, we take this opportunity to introduce two metrics for the content trait which are meta-content based—specificity of content and verbosity level. Texts that do not have good mix of general and specific content and writing which is verbose may be inefficient in conveying the content. These metrics are among the first to model content quality of articles in this manner.

The topic of an article also influences its quality. The area of article and book recommendation [108, 119] focuses on predicting topics which are interesting to an individual user. Here systems identify topics (approximated by words) in the set of articles which the user has already read. These topics are taken as indications of the type of content which is preferred by the user. New articles are suggested which are on similar topics as those which were identified. But there has been little work on understanding which topics are preferred in a larger readership, for example in a domain or genre.

In our analysis of science journalism quality, we are able to provide a preliminary analysis of some of the inherently interesting topics in this genre. We obtain a set of excellent writing samples from an anthology on best science writing and show that certain topics such as Medicine and Health, and Space appear to be typically more engaging to readers.

[8] http://www-nlpir.nist.gov/projects/duc/index.html

2.2.6 Genre and text quality

The texts selected for readability and coherence prediction tasks in prior work are also worthy of discussion.

Most readability measures were based on cognitive factors. As a result, these features are assumed as relevant for most texts [30, 49, 58]. A few studies have considered genre more explicitly. Some of them use common readability metrics or features on the new genre and do not focus on measures that may be unique to the genre under study. For example, Miltsakaki and Troutt (2007) [107] use readability formulae for web text and language models have been used for scientific articles [147] and web texts [25, 73]. Similarly, Zhao and Kan (2010) [173] propose a graph-based algorithm for measuring readability of concepts as well as documents mentioning the concepts within a domain. A readability score for each concept and each document is computed based on the idea that difficult concepts will be present in difficult documents and easier concepts in easier documents. They evaluate their method on medical and math domains. In contrast, there are studies with greater focus on a single genre. Elhadad (2006) [40] explores readability of technical articles from the medical domain and specifically seeks to identify medical terminology that would be unfamiliar to lay readers. Ma et al. (2012) [96] study readability prediction for children's books. They utilize features related to the visual layout of the book's pages such as font and image sizes which are rather specific to this genre.

A few data-driven organization metrics have also been evaluated on different genres. Barzilay and Lee (2004) [8] introduce a metric that tracks the subtopic structure of documents in a domain. It can be trained on texts from a given domain so that it is able to capture the topic transition properties for that domain. The basic idea of subtopic structure remains the same but can be adapted for any individual genre by training on that domain's texts. The focus of these experiments was to show the robustness of the metric rather than identify other aspects of coherence for the individual genres. In our work, the motivation for using genres is both for comparing the robustness of our metrics as well as explore a wider range of metrics by taking advantage of the distinctive properties of different genres.

2.3 Genres used in this thesis

This section provides an overview of the three genres which are the focus of this thesis. These genres vary in terms of writer competency. Automatic summaries can be expected to contain the most errors with regard to writing quality. Academic writers comprise a mix, novice, expert researchers and non-native speakers of the language. In the case of science journalism, the authors are professional and trained writers and so the average writing quality is very high. Therefore these texts are good and deficient in different matters of quality. Below we detail which differences (related to quality) in writing are most noticeable in these genres and how predicting text quality in these genres is useful for applications.

2.3.1 System generated summaries

The desired qualities of a summary are that it should contain important content from the source text, be concise and well-written. Automatic generation of coherent text is a hard problem. Therefore, most automatic summarization systems do not generate completely new sentences based on the source content. Rather they *extract* full sentences from the source document and use them to compose summaries. This approach leads to several problems in the quality of the output text.

Firstly, since extraction of full sentences is done, individual sentences can have content that is unnecessary given the context of the summary. Such unnecessary content could lead to verbose summaries. Sentence compression techniques [21, 52, 76, 103] aim to tackle this issue by shortening sentences while at the same time retaining important content and keeping the sentences grammatical. Further since the sentences come from different parts of a source document or from different documents (in the case of multi-document summarization), the ordering of sentences in a summary could also be problematic. Adjacent sentences in extractive summaries often lack continuity and to this end, a number of techniques have been developed to help re-organize summary content or score its organization quality [6, 80, 123]. In addition, referring expressions in the extracted sentences also need to be updated given the new context of these sentences in the summary. Some solutions proposed for this problem are automatic methods for revising

references to people [115] and also generic noun phrases [112] in summaries.

Some of these issues can be noticed in the example automatic summary (taken from the Document Understanding Conference datasets) below. The summary was produced from multiple source documents on the topic of a tunnel.

> Another is digging from the French coast in what the tunnel builders call the largest civil engineering project now under way in the world. LONDON – A shadow is falling across the light at the British end of the Channel Tunnel. While the French are forging ahead with a high-speed rail link to their end of the tunnel, for example, state-owned British Rail is dragging. Britain and France were linked beneath the English Channel on Tuesday when workers used a two-inch probe to connect two halves of a 31-mile undersea rail tunnel, officials reported. On May 6 the Queen and President Mitterrand will declare the Channel Tunnel open.

Therefore measures to evaluate the content and writing quality of summaries can have great use in development of summarization systems and their evaluation.

2.3.2 Academic writing

Clear writing is of great importance in academic publications. Consider the following abstract taken from a paper [60] which received the Best Paper Award at the HLT-NAACL conference in 2010.

> Coreference resolution is governed by syntactic, semantic, and discourse constraints. We present a generative, model-based approach in which each of these factors is modularly encapsulated and learned in a primarily unsupervised manner. Our semantic representation first hypothesizes an underlying set of latent entity types, which generate specific entities that in turn render individual mentions. By sharing lexical statistics at the level of abstract entity types, our model is able to substantially reduce semantic compatibility errors, resulting in the best results to date on the complete end-to-end coreference task.

The abstract has a clear structure going from introduction of the task and issues to the authors' ideas and results. Understanding what factors contribute to a good structure for

academic writing can help in at least two tasks.

There is little training for novice researchers, students and non-native speakers on how to write research papers and how to convey technical content in a clear manner. Automatic metrics to judge the quality of writing can provide feedback to such an audience. In addition, for novice students, it may also be helpful to suggest reading material such that the recommended papers not only convey the best ideas in the field but are also written well.

In addition to assessment, metrics for quality of academic writing would be quite useful for generation systems. Particularly there is interest in recent years on summarization of scientific articles [105, 130, 131]. So far, these studies have only focused on content selection from the papers. To create coherent and well-formed summaries text quality measures, also specific to this genre, are necessary.

There is fairly good spelling and grammar in most articles in this genre since some editing and review is done before publication. We expect that problems in organization and other discourse aspects and clarity of writing are the most relevant aspects to explore for quality prediction in this genre.

2.3.3 Science journalism

Science journalism has the most advanced writers. Here the authors are professional journalists and in our corpus which we describe later in this thesis, the articles for assessment are taken from the New York Times newspaper. So these articles have high quality in general. Moreover, while most news related to events focus on presenting facts, science journalism is meant to explain and also entertain. Consider the following snippet taken from an article by David Quammen and which appeared in the Harper's magazine.

> One morning early last winter a small item appeared in my local newspaper announcing the birth of an extraordinary animal. A team of researchers at Texas A&M University had succeeded in cloning a whitetail deer. Never done before. The fawn, known as Dewey, was developing normally and seemed to be healthy. He had no mother, just a surrogate who had carried his fetus to term. He had no father, just a "donor" of all his chromosomes. He was the genetic duplicate of a certain trophy buck out of south Texas whose skin cells had been

> cultured in a laboratory. One of those cells furnished a nucleus that, transplanted and rejiggered, became the DNA core of an egg cell, which became an embryo, which in time became Dewey. So he was wildlife, in a sense, and in another sense elaborately synthetic. This is the sort of news, quirky but epochal, that can cause a person with a mouthful of toast to pause and marvel. What a dumb idea, I marveled.

The passage provides much detail about the research. But it is also written in a clever story-like manner. Since writers in this genre employ many different techniques to create engaging articles, this genre presents the opportunity to examine which properties of writing are better at captivating reader interest. Among the six traits, we expect that there is most variation in the aspects related to creative writing style—the voice, sentence fluency and word choice traits.

Automatic measures to identify interesting articles can be useful in search and recommendation applications mainly. In addition, educational settings can also benefit from such metrics. For example, a high school teacher can use such metrics to select well-written science journalism articles to supplement readings from text books.

2.4 Gold-standards for text quality

The task of text quality prediction depends on strong and reliable judgements of which articles are of good and poor quality. Specifically, we focus on two needs for our corpora:

- that an expert reader is the audience of the text and that personal interests influence ratings only minimally.
- that the texts adequately capture distinctions for the trait we examine—content, organization or style. We should also be able to assume that grammar and other convention-related errors are nonexistent.

2.4.1 Automatic summarization

Out of the three genres in our work, the automatic summarization domain has the most direct text quality ratings available and also on a large scale. These ratings come from the

annual summarization workshops (DUC[9] and TAC[10]) organized by NIST. These workshops have been organized for over a decade now and evaluate summarization systems on a common test set each year. The output from systems are manually rated by NIST assessors for both their content (informative nature) as well as their linguistic quality (well-written nature). Linguistic quality ratings are done on a scale, for example, 1 to 10. On the other hand, content is evaluated by comparison with a human-written summary for the same source document. This evaluation practice has also created a wealth of human-written summaries, paired with their source documents.

The evaluation protocols are reliable and evolved over multiple years of research. The NIST assessors are retired information analysts and hence experts for the task of manual summary creation and quality judgements. Care is also taken to assign summaries to judges in such a way that the identity of the judge does not influence the resulting scores on the summaries.

Most automatic summarization systems are extractive and create a summary by adding complete sentences from the source article. Given this setup, and the source articles coming mostly from newswire, it is reasonable to assume that sentences mostly have correct grammar and spelling.

2.4.2 Academic writing

A corpus of ratings for academic writing quality would involve a much more sophisticated design. We would like to have ratings from experts (researchers) on the topic. But they are likely to be extremely influenced by their interests and familiarity with related work on the topic. Hence these ratings require a focused annotation design. For now, we consider this annotation task as beyond the scope of our work.

Rather, in our experiment we use an approach that is commonly followed in prior work on coherence modeling [7, 8, 41, 45, 71, 90]. We take an original text as an example of a well-organized article. Then we randomly permute its sentences and consider the permuted text as incoherent. For an expert reader, which we assume in our work, a permuted text would appear more incoherent compared to the original article. Recent

[9] http://duc.nist.gov

[10] http://www.nist.gov/tac/

work in Lin, Ng and Kan (2011) [90] show this hypothesis to be true: people when shown original and permuted versions of news articles, can tell the original apart with over 90% accuracy. Hence this dataset while simple, is functional and useful for validating our metrics in this area.

Our data comes from the ACL anthology corpus [135]. This resource contains the full text of all computational linguistics papers that were published at ACL (Association for Computatinal Linguistics) conferences starting from year 1965. Since these articles are reviewed before publication, it is reasonable to assume that the set of coherent and artificially created incoherent examples have few errors with regard to grammar and spelling. It has been observed that non-native speakers do make some errors in grammar while writing academic papers [31] but we believe that they are not so high to disrupt the evaluation of organization-related metrics.

2.5 Conclusions

In this chapter, we presented the setting for our work. We showed how educational rubrics provide an easy way to define the text quality problem in terms of aspects teachers notice and look for in student writing. We believe this framework will also help future work by other researchers to easily fit into particular traits. With this overall model of quality aspects, we also notice how prior work is mostly focused on predicting if the grammatical conventions of the language are followed and how the text is organized. Reader interest and content properties of texts have been little considered. In the following chapters, we fill this gap by proposing new metrics related to content, organization and interesting nature of articles. This chapter also described the existing resources for evaluation data for summarization and academic writing genres. We introduce a new corpus resource for studying text quality for science journalism. The next chapter presents the details of this corpus.

Chapter 3

A corpus of text quality for science journalism

For science journalism, we created our own corpus of text quality ratings. In this chapter, we describe the method by which we collected and categorized the articles in our corpus.

As we discussed in the chapters so far, the 'reader interest' aspect of quality has been little understood and very few computational methods exist which predict writing which is interesting and engaging to readers. In order to tackle this task, a suitable genre of articles should be selected where creative language use and interesting nature of writing is valued. Such language should also be frequently present in the texts in the genre so that we can study their relationship to quality. Fiction, poetry, and essay are some genres with such characteristics. Another apt genre is science journalism. Science news articles contain informative and at the same time entertaining content. This genre is in fact more balanced and reliable, compared to other genres, for studying text quality and analyzing properties of writing which contribute to reader interest. In genres such as fiction, personal tastes and preferred topics would have an enormous impact on quality perception and it would be difficult to focus on writing characteristics.

We therefore choose the science journalism genre for our study. In order to obtain articles with different quality levels, we use a simple heuristic of differentiating writing by renowned journalists from others. Our corpus contains several thousand articles, divided into three coarse levels of writing quality. All articles in the corpus have been published

in the New York Times (NYT), so the quality of any article is high. A small sample of GREAT articles was identified with the help of subjective expert judgements of established science writers. A substantially larger set of VERY GOOD writing was created by identifying articles published in the NYT and written by writers whose texts appeared in the GREAT section of the corpus. Finally, science-related pieces on topics similar to those covered in the GREAT and VERY GOOD articles but written by different authors formed the set of TYPICAL writing.

This corpus is suitable for studying text quality in a number of ways:

The corpus has both the desirable features which we discussed in the previous chapter—expert ratings and absence of conventions-related errors. The selection of the GREAT articles are made by expert and renowned journalists from an initial set of nominations and the VERY GOOD category is created by adding more samples from the authors of GREAT articles. We can also assume that all the categories of articles have excellent grammatical correctness and good organization since they are written by trained journalists and undergo review and edit procedures before they are published. In addition, all the articles come from the New York Times and therefore have high and reasonable quality on average. Therefore the corpus is suitable for analyzing aspects related to reader interest without consideration of lower level quality problems.

Our corpus is also a more realistic dataset of quality differences compared to prior studies. Previous work on quality prediction either used texts generated from automatic systems that have been rated for quality or they created poor quality examples by artificially manipulating an article. Text generated by automatic systems are quite different from those written by people. Again, manipulating articles to create negative samples is also far from realistic problems with quality. In contrast, our corpus contains reasonable text quality categories and is more relevant to the target applications of information retrieval and recommendation systems compared to previously used datasets.

The corpus is large scale containing thousands of articles. Therefore there is adequate data for training models with several features and varied test data for evaluation.

Below we provide specific details about the collection and categorization of articles in the corpus. We create two versions of our corpus. One contains labels corresponding to three levels of quality—GREAT, VERY GOOD and TYPICAL. A second corpus contains clusters of articles on the same topic, where each cluster has one very good writing sample and a set of 10 typical articles. The two corpora allow us to examine writing differences both across topics and within the same topic.

3.1 Creating general categories

All articles in our corpus were published in the New York Times between 1999 and 2007. By collecting all articles from the same source, we attempted to remove concerns about changes in article quality depending on the source of news.

3.1.1 Selecting GREAT articles

The GREAT articles in our corpus come from the "Best American Science Writing" annual anthologies. The stories that appear in these anthologies are chosen by prominent science journalists who serve as editors of the volume, with a different editor overseeing the selection each year. In some of the volumes, the editors explain the criteria they have applied for selecting articles:

> "First and most important, all are extremely well written. This sounds obvious, and it is, but for me it means the pieces impart genuine pleasure via the writers' choice of words and the rhythm of their phrases... "I wish I'd written that", was my own frequent reaction to these articles." (2004)

> "The best science writing is science writing that is cool... I like science writing to be clear and to be interesting to scientists and nonscientists alike. I like it to be smart. I like it, every once in a while, to be funny. I like science writing to have a beginning, middle and end—to tell a story whenever possible." (2006)

> "Three attributes make these stories not just great science but great journalism: a compelling story, not just a topic; extraordinary, often exclusive reporting; and a

facility for concisely expressing complex ideas and masses of information." (2008)

Therefore the articles in the "Best American Science Writing" anthologies present a wonderful opportunity to test computational models of structure, clarity, humor and creative language use.

We only select the articles in these anthologies which originally appeared in the New York Times newspaper. We limit the articles to this source because it is easier to select articles for the other text quality categories from the same newspaper. This selection is possible due to the availability of the New York Times Corpus [142] which contains the full text for NYT articles published for 20 years between 1987 to 2007. The NYT corpus also has extensive metadata including author information and editor assigned topic tags.

The Best Science Writing anthologies have been published since 1999 and the NYT corpus has articles up to year 2007. Therefore for this timespan it is straightforward to obtain the full text of the anthology articles from the NYT corpus. There are 63 articles which overlapped and they form the set of GREAT writing.

Obviously, the topic of an article will influence the extent to which it is perceived as well-written. We use the topic tags in NYT corpus metadata to provide a first characterization of the articles we got from the "Best American Science Writing" anthology. There are about 5 million unique tags in the full NYT corpus and most articles have five or six tags each. The number of unique tags for the set of GREAT writing articles is 199 which is too big to present. Instead, in Table 3.1 we present the tags that appear in more than three articles in the GREAT set. Medicine, space and physics are the most popular subjects in the collection. Computers, finance and mathematics topics are much lower in the list.

Next we describe the procedure we used to expand the corpus with samples of VERY GOOD and TYPICAL writing.

3.1.2 Extraction of VERY GOOD and TYPICAL writing

The number of GREAT articles is small—just 63—so we expanded the collection of good writing using the NYT corpus. The set of VERY GOOD writing contains NYT articles about research that were written by authors whose articles appeared in the GREAT sub-corpus. For the TYPICAL category, we pick other articles published around the same time but were

Tag	Articles	Tag	Articles
Medicine and Health	22	Computers and the Internet	4
Research	18	Doctors	4
Space	14	Drugs (Pharmaceuticals)	4
Science and Technology	13	Evolution	4
Physics	10	Planets	4
Biology and Biochemistry	8	Stem Cells	4
Genetics and Heredity	8	Age, Chronological	3
Archaeology and Anthropology	7	Brain	3
Reproduction (Biological)	7	Cloning	3
DNA (Deoxyribonucleic Acid)	6	Earth	3
Animals	5	History	3
Diseases and Conditions	5	Mental Health and Disorders	3
Ethics	5	Religion and Churches	3
Finances	5	Universe	3
Women	5	Vaccination and Immunization	3

Table 3.1: Most frequent topic tags in the GREAT writing samples

neither chosen as best writing nor written by the authors whose articles were chosen for the anthologies. We followed two steps to create the VERY GOOD and TYPICAL categories.

Finding a relevant set

The NYT corpus contains every article published between 1987 to 2007 comprising a few million articles in total. We first filter some of the articles based on topic and research content before sampling for good and typical examples. The goal of the filtering is to find articles about science that were published around the same time as our GREAT samples and have similar length. We consider only:

- Articles published between 1999 and 2007. This is the period for which the best science writing anthologies have been published.
- Articles that are at least 500 words long. All articles from the anthologies had that minimum length.
- Only science journalism pieces.

In the NYT metadata, there is no specific tag that identifies all the science journalism articles. So, we create a set of metadata tags which can represent this genre. Since we know the GREAT article set to be science writing, we choose the minimal subset of tags such that at least one tag per GREAT article appears on the list. We call this set as "science tags". We derived this list using greedy selection, choosing the tag that describes the largest number of GREAT articles, then the tag that appears in most of the remaining articles, and so on until we obtain a list of tags that covers all GREAT articles. Table 3.2 lists the fourteen topic tags that made it into the "science tags" list.

We consider an article to be science related if it has one of the topic tags in "science tags" and also mentions words related to science such as 'scientist', 'discover', 'found', 'physics', 'publication', 'study'. We found the need to check for words that appear in the article because in the NYT, research-related tags are assigned even to articles that only cursorily mention a research problem such as stem cells but otherwise report general news. We used a hand built dictionary of research words and remove articles that do not meet a threshold for research word content. The dictionary comprises a total of 73 lexical

Medicine and Health	22
Space	14
Research	8
Physics	4
Evolution	3
Computers and the Internet	2
Religion and Churches	2
Language and Languages	2
Biology and Biochemistry	1
Animals	1
Brain	1
Light	1
Global Warming	1
Baseball	1

Table 3.2: Minimum set of "science tags" which cover all GREAT articles. The tags are listed in the order in which they were selected by greedy approach. The count indicates the number of articles covered by the tag during the selection process. The 'Medicine and Health' tag covers 22 articles and 'Space' covers 14 of the *remaining* articles and so on.

People	Process	Topic	Publications	Endings	Other
researcher	discover	biology	report	-ology	human
scientist	found	physics	published	-gist	science
physicist	experiment	chemistry	journal	-list	research
biologist	work	anthropology	paper	-mist	knowledge
economist	finding	primatology	author	-uist	university
anthropologist	study		issue	-phy	laboratory
environmentalist	question				lab
linguist	project				
professor	discuss				
dr					
student					

Table 3.3: Unique words from the research word dictionary

items including morphological variants. Six of the entries in the dictionary are regular expression patterns that match endings such as "-ology" and "-gist" that often indicate research related words. The unique words from our list are given in Table 3.3. We have grouped them into some simple categories here.

An article was filtered when (a) fewer than 10 of its tokens matched any entry in the dictionary or (b) there were fewer than 5 unique words from the article that had dictionary matches. This threshold keeps articles that have high frequency of research words and also diversity in these words. The threshold values were tuned such that all the articles in the GREAT set, scored above the cutoff. After this step, the final *relevant* set has 23,710 science-related articles on the same topics as the GREAT samples.

Subdividing the relevant set

The GREAT articles were written by 40 different authors. Some authors have more than one article appearing in that set, and a few have even three or more articles in that category. The top 10 authors according to the number of their articles in the GREAT set are listed in Table 3.4.

It is reasonable to consider that the writers of the GREAT samples are exceptionally

Author	No. of articles in GREAT set
Dennis Overbye	9
Natalie Angier	9
Nicholas Wade	3
Gardiner Harris	3
Stephen S. Hall	3
Alan Lightman	2
Daniel C. Dennett	2
Janet Roberts	2
Lawrence K. Altman	2
William J. Broad	2

Table 3.4: Top authors in the GREAT writing set

Category	No. articles	No. sentences	No. tokens
GREAT	63	7,212	177,775
VERY GOOD	4,190	232,824	5,924,189
TYPICAL	19,520	1,213,534	30,152,575
Total	23,773	1,453,570	36,254,539

Table 3.5: Overview of GREAT, VERY GOOD and TYPICAL categories in the corpus

good, so we extracted all articles from the *relevant set* written by these authors (all 40 authors) to form the VERY GOOD set. There are 4190 in that category. The remaining articles from the *relevant set*, 19520 in number, are grouped to form the TYPICAL class.

A summary of the three categories of articles is given in Table 3.5.

3.2 Topic-normalized corpus

As we already noted in the previous section, the articles in our corpus span a wide variety of topics. The writing style for articles from different topics, for example, health vs. religion research would be widely different and hard to analyze for quality differences. In addition, during information retrieval, one would need to compare topically similar

(relevant to the query) articles. So we create another corpus which contains clusters of topically related articles derived from the general categories we obtained above.

For each article in the GREAT and VERY GOOD sets, we associate a list of articles from the TYPICAL category which discuss the same or closely related topic. To identify topically similar articles, we compute similarity between the articles. Only the descriptive topic words identified via a log likelihood ratio test are used in the computation of similarity. The descriptive words are computed using the TopicS tool[11] [94]. Each article is represented by binary features which indicate the presence of each topic word. The similarity between two articles is computed as the cosine similarity between their vectors.

For each GREAT and VERY GOOD article, we store the list of 10 most similar TYPICAL articles. According to our observations, these article clusters are highly coherent with regard to topic and discuss very closely related content. The mappings for two GREAT/VERY GOOD articles are demonstrated in Table 3.6 by listing the titles of the articles.

The chosen 10 most similar TYPICAL articles for each GREAT or VERY GOOD article actually have varied similarity values (from 0.06 to 0.6). We will use these values in later chapters to examine how the accuracy of text quality prediction varies with different levels of topic normalization. Tables 3.7 and 3.8 give a further example of a GREAT article and one of the matched TYPICAL article. These articles have a similarity value of 0.28. A snippet from the beginning of each article is shown.

In this way, for many of the high quality articles we have collected examples with hypothesized inferior quality, but on the same topic.

To create the data for classification experiments in this thesis, we pair up each GREAT and VERY GOOD articles with each of the TYPICAL articles in the similar articles list. The total pairs created is 42530. The task is to identify the VERY GOOD or GREAT article in the pair.

3.3 Analysis of author bias

For the creation of our categories we assumed that writers whose articles were published in the "Best American Science Writing" are great authors in general. All their articles were

[11] http://www.cis.upenn.edu/~lannie/topicS.html

VERY GOOD **or** GREAT **article:** *Human Genome May Be Longer Than Expected* **Matched** TYPICAL **articles:** READING THE BOOK OF LIFE: What Lies Ahead; Journey to the Genome Huge Genome Project Is Proposed to Fight Cancer A New Kind of Genomics, With an Eye on Ecosystems 50,000 Genes, and We Know Them All (Almost) A human gene is patented as a potential tool against AIDS, but ethical questions remain. Genome Pioneer Will Start Center of His Own A DNA Chip Maker Acquires Gene-Sequencing Company Agriculture Takes Its Turn in the Genome Spotlight Citing RNA, Studies Suggest A Much Deeper Gene Pool Speed-Reading the Book of Life
VERY GOOD **or** GREAT **article:** *Quantum Stew: How Physicists Are Redefining Reality's Rules* **Matched** TYPICAL **articles:** How Does a Photon Decide Where to Go? That's the Quantum Mystery One Hundred Years of Uncertainty Quantum Weirdness Physics' Big Puzzle Has Big Question: What Is Time? Space-Time Is of the Essence The Universe on a String 3 Researchers Based in U.S. Win Nobel Prize in Physics Where Protons Will Play Quantum Leap May Transform Chips The Story of H

Table 3.6: Two example clusters of GREAT or VERY GOOD article paired with 10 most similar TYPICAL articles

GREAT **writing sample**

Kristen Ehresmann, a Minnesota Department of Health official, had just told a State Senate hearing that vaccines with microscopic amounts of mercury were safe. Libby Rupp, a mother of a 3-year-old girl with autism, was incredulous. “How did my daughter get so much mercury in her?” Ms. Rupp asked Ms. Ehresmann after her testimony. “Fish?” Ms. Ehresmann suggested. ”She never eats it,” Ms. Rupp answered. “Do you drink tap water?” “It’s all filtered.” “Well, do you breathe the air?” Ms. Ehresmann asked, with a resigned smile. Several parents looked angrily at Ms. Ehresmann, who left. Ms. Rupp remained, shaking with anger. That anyone could defend mercury in vaccines, she said, ”makes my blood boil.” Public health officials like Ms. Ehresmann, who herself has a son with autism, have been trying for years to convince parents like Ms. Rupp that there is no link between thimerosal—a mercury-containing preservative once used routinely in vaccines – and autism. They have failed. The Centers for Disease Control and Prevention, the Food and Drug Administration, the Institute of Medicine, the World Health Organization and the American Academy of Pediatrics have all largely dismissed the notion that thimerosal causes or contributes to autism. Five major studies have found no link. Yet despite all evidence to the contrary, the number of parents who blame thimerosal for their children’s autism has only increased. And in recent months, these parents have used their numbers, their passion and their organizing skills to become a potent national force. The issue has become one of the most fractious and divisive in pediatric medicine.

Table 3.7: Snippet from a GREAT article

TYPICAL **writing sample**

Neal Halsey's life was dedicated to promoting vaccination. In June 1999, the Johns Hopkins pediatrician and scholar had completed a decade of service on the influential committees that decide which inoculations will be jabbed into the arms and thighs and buttocks of eight million American children each year. At the urging of Halsey and others, the number of vaccines mandated for children under 2 in the 90's soared to 20, from 8. Kids were healthier for it, according to him. These simple, safe injections against hepatitis B and germs like haemophilus bacteria would help thousands grow up free of diseases like meningitis and liver cancer. Halsey's view, however, was not shared by a footnotesize but vocal faction of parents who questioned whether all these shots did more harm than good. While many of the childhood infections that vaccines were designed to prevent – among them diphtheria, mumps, chickenpox and polio – seemed to be either antique or innocuous, serious chronic diseases like asthma, juvenile diabetes and autism were on the rise. And on the Internet, especially, a growing number of self-styled health activists blamed vaccines for these increases.

Table 3.8: Snippet from a TYPICAL article which is topically related to the GREAT article in Table 3.7

chosen into the VERY GOOD writing set. Articles written by other authors were collected as TYPICAL writing.

However, some authors may write a greater volume of articles compared to others. Therefore one concern about the corpus is that the categories could be skewed towards articles from only a few authors. In this section, we analyze how many articles from different authors are present in each of our categories and the topic normalized corpus.

Table 3.9 shows the top authors according to number of their articles present in a category. The details are provided for two categories—the set of GREAT and VERY GOOD articles combined, and the set of TYPICAL articles. Since the GREAT and VERY GOOD set is based primarily on 40 authors (the author list in the expanded set of VERY GOOD articles is larger than 40 since some articles are co-authored by more than one person), the proportion of articles coming from individual authors is high. The top 4 authors in this category each have close to 10% share of the articles. However, the dataset is not very skewed. To cover 85% of the articles, around 15 authors are necessary. Therefore the category is not indicative of the writing style of two or three authors only. For the TYPICAL category, the proportion of articles from any individual author is even lower, the highest is 2.4% from one author. The top 15 authors contribute to only 16% of the category.

Similarly, we computed the most frequent pairs of authors for the (GREAT or VERY GOOD, TYPICAL) article pairs in our topic normalized corpus. These results are presented in Table 3.10. The most frequent author pair comprises close to 942 pairs (2.2%) out of the 42530 pairs in our corpus. The top 15 author pairs make up 13% of the data. Therefore the topic normalized corpus also has considerable variety in the pairing of articles by different authors. So we can expect to learn general writing differences across the set of very good and average writers rather than the writing style of individual authors.

3.4 Comparison with ratings of a student annotator

The seed articles for our corpus were obtained using the judgements of leading journalists. For expanding this set, we used a simple heuristic based on the authors of these seed articles. Therefore we can consider the resulting categories as approximating the judgements of the expert journalists. In this section, we provide the results of a small an-

VERY GOOD +GREAT **writing**		TYPICAL **writing**	
Author	**No. (%) of articles**	**Author**	**No. (%) of articles**
Altman, Lawrence K	417 (9.8)	Fountain, Henry	466 (2.4)
Kolata, Gina	407 (9.6)	Pollack, Andrew	380 (1.9)
Wade, Nicholas	371 (8.7)	Markoff, John	306 (1.6)
Grady, Denise	354 (8.3)	Lohr, Steve	280 (1.4)
Chang, Kenneth	298 (7.0)	Revkin, Andrew C	213 (1.1)
Brody, Jane E	273 (6.4)	Schwartz, John	209 (1.1)
Wilford, John Noble	254 (6.0)	Pear, Robert	183 (0.9)
Stolberg, Sheryl Gay	253 (5.9)	Leary, Warren E	167 (0.8)
Mcneil, Donald G Jr	170 (4.0)	Glanz, James	165 (0.8)
Overbye, Dennis	166 (3.9)	Goode, Erica	160 (0.8)
Broad, William J	157 (3.7)	Goodstein, Laurie	146 (0.7)
Harris, Gardiner	140 (3.3)	Blakeslee, Sandra	133 (0.7)
Carey, Benedict	132 (3.1)	Feder, Barnaby J	132 (0.7)
Harmon, Amy	122 (2.9)	Hafner, Katie	132 (0.7)
Gorman, James	100 (2.4)	Eisenberg, Anne	130 (0.7)
	3614(85.0)		3202 (16.3)

Table 3.9: The 15 most frequent authors in the GOOD and TYPICAL categories

Author pair		No. (%) of examples
Author of VERY GOOD article	**Author of TYPICAL article**	
Wade, Nicholas	Pollack, Andrew	942 (2.2)
Overbye, Dennis	Glanz, James	622 (1.5)
Wilford, John Noble	Leary, Warren E	516 (1.2)
Carey, Benedict	Goode, Erica	375 (0.9)
Chang, Kenneth	Leary, Warren E	372 (0.9)
Chang, Kenneth	Glanz, James	346 (0.8)
Kolata, Gina	Pollack, Andrew	324 (0.8)
Wilford, John Noble	Glanz, James	323 (0.8)
Altman, Lawrence K	Pollack, Andrew	320 (0.8)
Grady, Denise	Pollack, Andrew	266 (0.6)
Altman, Lawrence K	Bradsher, Keith	220 (0.5)
Kolata, Gina	Duenwald, Mary	218 (0.5)
Chang, Kenneth	Fountain, Henry	211 (0.5)
Overbye, Dennis	Leary, Warren E	207 (0.5)
Grady, Denise	Duenwald, Mary	195 (0.5)
	Total	5457 (13%)

Table 3.10: The 15 most frequent author pairs of VERY GOOD and TYPICAL articles in the topic normalized corpus

notation study where we asked an undergraduate student to provide personalized ratings for a few articles from our corpus. We wanted to study the following questions:

1. How much does an individual's ratings agree with the experts?
2. Is there noticeable difference between the GREAT and VERY GOOD categories?
3. How accurate is the similarity measure used for creating the article mappings in the topic normalized corpus?

As we discussed in Chapter 2, in this thesis, we use the ratings of experts as our gold standard because that definition helps us focus on the linguistic properties of the text. We performed the following annotation study in order to understand how a person from the target population of an application (such as a recommendation system) would rate the same articles. People differ in which topics they like and have personal preferences for style of writing. It is hard to control for these preferences during annotation. However it is useful to know how the judgements of a target population relates to expert ratings which we use for developing the text quality measures. This annotation study is a preliminary analysis with this aim.

We hired an undergraduate student to do the annotations. The student had no prior knowledge and experience in natural language processing techniques or linguistics. From our topic normalized corpus we chose 20 pairs of (GREAT, TYPICAL) articles and 20 pairs of (VERY GOOD, TYPICAL) articles for annotation. We also created 10 pairs, where both articles came from the GREAT or VERY GOOD categories. In each case, the TYPICAL article is one of the 10 most similar articles to the GOOD sample but they span a range of similarity values as noted in the previous section.

The student read each article in a pair and answered two questions. The order of articles in a pair was randomly assigned and the pairs were also randomly presented. A computer interface was used for the annotation. It showed the two articles on the screen and the following questions.

Is the topic of the articles the same? For example, when both articles are about 'controversies related to vaccination' we may consider them highly similar. When

both are about 'vaccines', they are medium similar and when one is about archaeology and other about chemistry, you may consider them not at all similar. Therefore varying degrees of similarity can be assigned to articles. The scale for this rating is 1 (not same) to 10 (almost exactly same).

Which article is more interesting to read? Give an overall rating for how much you would prefer to read one article versus another. You may find one article more interesting because it is more informative, written creatively or captivates your attention. Indicate your preference on the following scale: a) prefer article A very much b) prefer article A somewhat c) no preference d) prefer article B somewhat e) prefer article B very much

We provided the annotator with 10 practice pairs of articles to familiarize herself with the task and scales for ratings. Then the 50 pairs that we described above were provided.

First, we provide an analysis of the similarity ratings from the annotator. We compare the automatic measure we used for pairing articles (cosine overlap of topic words) with the annotator's ratings for similarity. These values are plotted in Figure 3.1. The Pearson correlation between the automatic measure and annotator scores is rather high, 0.57 (pvalue of 1.5e-5). Therefore the similarity metric used for topic normalization is quite reliable.

For the ratings of quality, we have summarized the results in Table 3.11. The first column indicates what type of pair was compared. The 'GOOD is better' column presents the number of examples where the annotator chose the GREAT or VERY GOOD article as better than the TYPICAL article. We had two levels of preference–'very much better' and 'better'. We present the combined counts for both these levels since the number of examples in our annotation study is not large. Similarity we indicate the number of times a TYPICAL article was preferred over the GREAT or VERY GOOD articles. 'No pref.' indicates that neither article was preferred over the other.

For the pairs comparing GREAT with a TYPICAL article we find that the GREAT article is chosen as better in 14 out of 20 pairs. This result indicates that the annotator had a clear preference for the GREAT articles, aligning with the judgements of the expert journalists.

The trend for the VERY GOOD versus TYPICAL articles is not as strong. Close to half the pairs were judged as 'no preference' and the remaining cases were almost equally divided

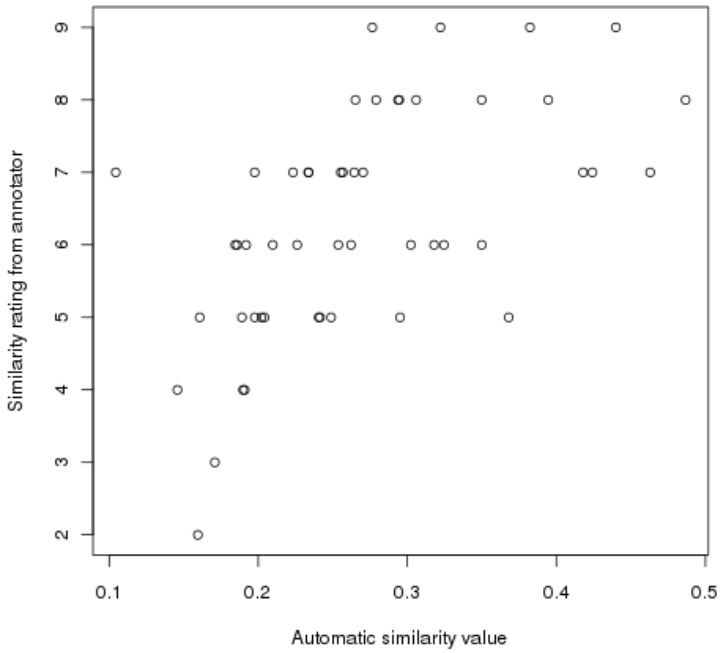

Figure 3.1: Similarity values computed using topic words versus annotator's similarity ratings

Type of pair	**No. pairs**	**No pref.**	GOOD **is better**	TYPICAL **is better**
GREAT VS. TYPICAL	20	1	14	5
VERY GOOD VS. TYPICAL	20	9	6	5
GREAT VS. VERY GOOD	10	5	3	2
Total pairs	50			

Table 3.11: Summary of quality ratings from the student annotator

between preferring the VERY GOOD (6 times) and TYPICAL (5 times) articles. Our simple heuristic of using the articles written by the GREAT author set as VERY GOOD writing is not reflected in these ratings. Further examination should be done to understand if our heuristic works well. For example, these ratings were provided by a student. It would be interesting to examine how a professional writer or journalism student will rate the same articles. Further obtaining ratings from a number of annotators and averaging them will provide better normalization over people's individual preferences. For now we will continue to use the categories developed by our heuristics and leave further annotation and cleaning of article categories for future work. For comparing GREAT and VERY GOOD articles however, the results are close to expected. Half the pairs are rated as 'no preference' indicating that both articles could be of good quality.

3.5 Setup for classification tasks

We perform two types of classification tasks in this thesis. We divide our corpus into development and test sets for these tasks in the following way.

Any topic: Here the goal is to separate out VERY GOOD (or GREAT) versus TYPICAL articles without regard to topic. The test set contains 4,153 VERY GOOD or GREAT articles and we randomly sample 4,153 articles from the TYPICAL category to comprise the negative set.

Same topic: Here we use the topic-paired VERY GOOD (or GREAT) and TYPICAL articles. The goal is to predict which article in the pair is the VERY GOOD or GREAT one. For test set, we selected 41,530 pairs.

Development data: We randomly selected 100 VERY GOOD articles and their paired (10 each) TYPICAL articles from the topic-normalized corpus. Overall, these articles constitute 1,000 pairs which we use for developing the same-topic classifier. From these selected pairs we take the 100 VERY GOOD articles and sample 100 unique articles from the TYPICAL articles making up these pairs. These 200 articles are used to tune the any-topic classifier. The two test sets above do not overlap with these development sets.

3.6 Conclusions

In this chapter, we described how using simple heuristics we were able to create a corpus with text quality categories for the science journalism genre. In later chapters, we provide text quality experiments on this corpus. We apply the organization model based on intentional structure to this genre in Chapter 4 and use specificity scores for doing the prediction in Chapter 5. Chapter 6 is fully devoted to text quality prediction specifically for science journalism. In that chapter, we develop several measures related to genre-specific aspects of science news and reader interest and examine the accuracy of predicting the quality categories which we developed here.

Apart from text quality assessment, given the widespread use of creative language in science journalism, our corpus could be useful for computational study of several language phenomenon. Particularly, there is frequent use of metaphor, figurative and humourous language in these articles. Within computational linguistics, there is a lot of interest in developing automatic methods to identify such language [9, 46, 106, 146]. Articles from our corpus can provide a good dataset for annotation of such constructions and developing methods for automatic detection. Further, good accuracies on these tasks have the potential to improve text quality prediction by allowing us to examine how metaphor, humor and idiomatic language are associated with reader perception of interesting nature and quality.

Chapter 4

A model of organization based on intentional structure

The order, presentation, and structure of the piece are compelling and guide the reader purposefully through the text.

Transitions are smooth, helpful and natural.

[Organization trait (Section 2.1)]

As we discussed in the related work sections, there are a number of metrics to score organization quality which were developed in previous work. We introduce a new measure based on the intentional structure of writing which we propose as well-suited for text quality analysis of academic writing and science journalism. This chapter explains the design and implementation of our metric.

While writing any article, an author has a purpose that he wishes to convey to the reader. For example, the purpose could be narrating an event, explaining a concept, critiquing an idea or supporting an argument. Discourse theories such as Grosz and Sidner (1986) [57] consider 'purpose' as a crucial component of discourse structure and as important for the discourse to be perceived as coherent. This theory also associates individual discourse segments in the article with 'intentions' that contribute towards achieving the

1a) An aqueduct is a water supply or navigable channel constructed to convey water. b) In modern engineering, the term is used for any system of pipes, canals, tunnels, and other structures used for this purpose.
2a) Cytokine receptors are receptors that bind cytokines. b) In recent years, the cytokine receptors have come to demand more attention because their deficiency has now been directly linked to certain debilitating immunodeficiency states.

Table 4.1: The first two sentences of two descriptive articles

overall purpose of the text. To this end, the discourse segments are connected by relations which, for example, indicate if a particular intention should be satisfied before another one.

For an example, let us consider the opening sentences of two descriptive articles[12] shown in Table 4.1. The purpose of these articles is to describe the concept of an aqueduct and a cytokine receptor. The first sentence of both these articles, sentences (1a) and (2a), are definitions. Their second sentences, (1b) and (2b), provide further specific details about the concept. Such regularity in intention sequences exists across articles with the same purpose and the familiar structure facilitates the reader's understanding of the subject matter.

In fact, in the academic writing genre, there are systematic studies of the intentional structure of articles. The reason for this interest is the argumentative nature of academic writing where researchers have the purpose of convincing the reader of the problems with prior approaches and the merits of their own solution. Moreover, since journal and conference publications represent a restrictive type of writing, these articles typically involve a small set of frequent intentions for their sentences. One of the first work in this area was done by Swales [154] who proposed that academic articles have three coarse segments. In the first segment called '*creating a territory*', authors describe motivation for

[12]Wikipedia articles on "Aqueduct" (`http://en.wikipedia.org/wiki/Aqueduct`) and "Cytokine Receptors" (`http://en.wikipedia.org/wiki/Cytokine_receptor`)

a problem. The second, '*establishing a niche*' puts forth the goal of the current research by identifying a gap in prior work or raising a question that needs to be solved. The final segment '*occupying the niche*' involves description of the new work and associated details.

Apart from theoretical studies, large scale annotations of intentional structure have been carried out on academic articles [85, 160, 162]. These annotations are done on sentence level and involve a small set of categories. For example, Teufel, Carletta and Moens (1999) [160] use 7 labels called *argumentative zones*, for example, *aim*, *contrast*, *basis* and *background*. Supervised classifiers have also been built to identify such categories on unlabelled data [59, 162]. Subsequently, these distinctions are being used in applications such as summarization of scientific papers [163] and for automatically tagging citation sentences with their function (criticism, basis, etc.) in the paper [164].

Although much understanding of the theory of intentional structure was gained by prior work, this idea has not been used in computational methods to predict the coherence of articles. But given its importance and particular relevance to research writing, intentional structure is an attractive model of organization for our work. Apart from academic writing, we also expect such a model to be relevant for the science journalism articles although the latter has a looser structure and the set of intentions could be much more varied than in academic articles. To this end, we developed a method to rate organization quality using the intentional structure of articles. In our approach, we learn the regularities in the intentional structure of well-written articles (having the same 'purpose') and use it to predict if a new test article conforms to the patterns of well-written structure.

We wanted our approach to be applicable for different genres particularly for science news which is similar to academic writing with regard to intentional structure. However, rather than create manual annotations of intentional structure to obtain training data, which would change for each genre, we make a simplifying assumption in our work. We assume that syntactic patterns can provide indications of intentional structure and sentences that have high syntactic similarity could be similar in intentions. We detail this idea in Section 4.1. This idea also requires us to examine how the syntax of sentences should be represented and how patterns over these should be defined. In Section 4.2, we

present the two methods that we used for representing syntax and Section 4.3 describes two approaches for using syntactic regularities to produce a score for organization quality.

Finally, we present evaluations of text quality in the two relevant genres—academic writing (Section 4.4) and science news (Section 4.5). In both cases, we provide comparison with other measures from prior work to predict organization quality. We find that our syntax model provides good performance above the baseline and comparable accuracies with other approaches for both these genres. On the academic genre, we also examine how manual annotations and supervised systems for intentional structure that were developed in prior work perform for this task of predicting organization quality. Using these annotations, we analyze whether some of our syntactic patterns correlate with intentional structure categories defined in previous theories.

The text quality evaluations on the academic genre are based on permutations based examples which we introduced in Section 2.4. An original article is taken as a well-organized article and a random permutation of its sentences is taken as an article with poor organization. We use these examples to validate that our model is useful for distinguishing differences in organization quality. Then we use the science journalism genre with its realistic corpus to further evaluate the performance of our method for text quality prediction.

4.1 Syntax as a rough proxy

In order to learn patterns in intentional structure, we need a way to identify the author intention behind each sentence. One approach would be to obtain annotations for intentions for a reasonable amount of data and use it to train a classifier to identify intentions. But such annotations are only available (publicly) on one corpus of chemistry academic journal articles [84] and more recently on a corpus of computational linguistics conference publications [158]. To perform the task on a different genre or even different subgenre of academic articles such as review summaries, we would need to obtain separate annotations. Further, annotation for intentional structure involves several challenges. For many genres, even the related area of science journalism, it would be challenging to predefine the intention categories and obtain reliable annotations. Academic articles have a

restricted structure and further can be analyzed as individual sections, but it is unclear if a similar strategy can be developed for other genres. Therefore rather than rely on manual annotation, we use an insight about sentence syntax to propose an approximate indicator of sentence intention.

We introduce the idea that the syntax of a sentence can act as a rough proxy for its intentional structure. The motivation for using syntax comes from the observation that certain sentence types such as questions and definitions have distinguishable and unique syntactic structure.

For instance, in our previously introduced example from Wikipedia articles (Table 4.1), several syntactic patterns can be found. The first sentences of these articles have the prototypical syntax of definition sentences. Definitions usually have the same structure: they start with concept to be defined expressed as a noun phrase followed by a copular verb (is/are). The predicate contains two parts: the first is a noun phrase reporting the concept as part of a larger class (eg. an aqueduct is a water supply), the second component is a relative clause listing unique properties of the concept. The second sentences of the articles (1b and 2b), which provide specific details also have some distinguishing syntactic features such as the presence of a topicalized phrase providing the focus of the sentence. In this way, the two articles which have similar sequence of communicative goals also have similar syntactic patterns for the sentences in the sequence.

A number of recent studies also support the idea of syntactic patterns in discourse. Cocco et al. (2011) [22] show that significant associations exist between certain part of speech tags and sentence types such as explanation, dialog and argumentation in French short stories. For the task of discourse parsing, Lin et. al (2009) [89] report that the syntactic productions from adjacent sentences are powerful features for predicting which discourse relation (cause, contrast, etc.) holds between them.

There is also evidence from entrainment literature that certain grammatical productions are repeated in adjacent sentences more often than would be expected by chance [38, 140]. Motivated by such patterns, Debey, Keller and Sturt (2006) [37] and Cheung and Gerald (2010) [20] build parsers that take advantage of the syntax of adjacent sentences for parsing a current sentence. The idea is that a production that was used in the

immediately previous sentence is likely to be relevant for the current sentence as well given the evidence from syntactic entrainment.

However, these entrainment-based studies have focused only on the repetition of grammatical productions in adjacent sentences. We performed a pilot study to examine if other types of syntactic patterns are also present in adjacent sentences. In this study we considered all pairs of grammatical productions and investigated whether they are likely to appear in adjacent sentences more often than chance.

We use the gold standard parse trees from the Penn Treebank [100] for this study. Our unit of analysis is a pair of adjacent sentences (S_1, S_2) and we choose to use Section 0 of the corpus which has 99 documents and 1727 sentence pairs. We enumerate all productions that appear in the syntactic parse of any sentence and exclude those that appear less than 25 times, resulting in a list of 197 unique productions. Then all ordered pairs[13] (p_1, p_2) of productions are formed. There are a total of 38,809 production pairs.

For each pair, we compute a 2x2 contingency table with the following components:

- $c(p_1 p_2) =$ number of sentence pairs where $p_1 \in S_1$ and $p_2 \in S_2$
- $c(p_1 \neg p_2) =$ number of pairs where $p_1 \in S_1$ and $p_2 \notin S_2$
- $c(\neg p_1 p_2) =$ number of pairs where $p_1 \notin S_1$ and $p_2 \in S_2$
- $c(\neg p_1 \neg p_2) =$ number of pairs where $p_1 \notin S_1$ and $p_2 \notin S_2$

We remove the pairs where $c(p_1 p_2)$ is less than three. Then we use a chi-square test to understand if the observed count $c(p_1 p_2)$ is significantly (95% confidence level) greater or lesser than the expected value if occurrences of p_1 and p_2 were independent.

Given that we are performing the tests for a large number of production pairs (38,809), there is an increased chance of Type I errors (rejecting the null hypothesis when it is actually true). To mitigate this issue, we perform Bonferroni correction for the p-values from the test. To ensure that an overall 95% confidence level is maintained (for the full set of tests), individual p-values should be less than $0.05/38809 = 1.28 \times 10^{-6}$ This approach is one of the conservative techniques to reduce Type I errors.

[13] (p_1, p_2) and (p_2, p_1) are considered as different pairs.

For this corrected p-value, 25 production pairs turn out as occurring significantly greater than chance. No pair was detected as occurring less than expected. The 25 pairs of the first kind are listed in Table 4.2 along with the number of times they occurred together, $c(p_1 p_2)$. We also divide these pairs into three simple categories: 'repetitions', 'related to quantities' and 'other'. In Dubey, Sturt and Keller (2005) and Cheung and Penn (2010), a similar test was peformed for identifying production pairs that are repeated very often in adjacent sentences. They use a slightly different test which examines if the probability with which the production appears in a second sentence S_2 given that it appeared in previous sentence S_1 is greater than the probability with which it generally appears in S_2. Cheung and Penn compute these productions also on the Penn Treebank albeit on different sections compared to our analysis. However, we present some of their results also for comparison. In the last column in Table 4.2, we show the top 10 productions which Cheung and Penn report in their paper as having the highest entrainment. Their list is weighted by the frequency of the production.

A small fraction of the significant pairs (7/25) that we found are indeed repetitions as pointed out by prior work. Most of these are related to quantifier phrases and noun phrases similar to the top list of Cheung and Penn. However, we also found other regularities which are not repetition of productions. Some of these sequences are related to quantities and can be explained by the fact that these articles come from the finance domain and often discuss prices and shares. But there is also a class that is not repetitions or readily observed as domain-specific.

We analyzed example sentences with these sequence patterns to understand some of the trends. The most frequent pattern, (VP $\rightarrow$ VB VP | NP-SBJ $\rightarrow$ NNP NNP), contains a bare verb in S_1 and propernames as subjects of the second. We found that in such sentence pairs, S_1 is often associated with modals and presents hypotheses or speculations. The following sentence S_2 often has an entity, a person or organization, giving their opinion on the hypothesis. This pattern roughly correponds to a SPECULATE followed by ENDORSE sequence of intentions in the sentences. An example sentence pair with these productions is shown below. The spans corresponding to the left-hand side non-terminal in the productions is indicated by square brackets.

Our study			Cheung and Penn (2010)
p_1	p_2	$c(p_1p_2)$	
REPETITIONS			
VP→VBD SBAR	VP→VBD SBAR	83	QP→# CD CD
QP→$ CD CD	QP→$ CD CD	18	NP→JJ NNPS
NP→$ CD -NONE-	NP→$ CD -NONE-	16	NP→NP , ADVP
NP→QP -NONE-	NP→QP -NONE-	15	NP→DT JJ CD NN
NP-ADV→DT NN	NP-ADV→DT NN	10	PP→IN NP NP
NP-LOC→NP , NP	NP-LOC→NP , NP	3	QP→IN $ CD
NP→NP NP-ADV	NP→NP NP-ADV	7	NP→NP : NP
RELATED TO QUANTITIES			INTJ→UH
NP→QP -NONE-	QP→$ CD CD	16	ADVP→IN NP
QP→$ CD CD	NP→QP -NONE-	15	NP→CD CD
NP→NP NP-ADV	NP→QP -NONE-	11	
NP-ADV→DT NN	NP→QP -NONE-	11	
NP→NP NP-ADV	NP-ADV→DT NN	9	
NP-ADV→DT NN	NP→NP NP-ADV	8	
NP-ADV→DT NN	NP→$ CD -NONE-	8	
NP→$ CD -NONE-	NP-ADV→DT NN	8	
NP→NP NP-ADV	QP→CD CD	6	
QP→CD CD	NP→NP NP-ADV	5	
FRAG→NP-SBJ NP	NP→$ CD -NONE-	3	
OTHER			
VP→VB VP	NP-SBJ→NNP NNP	27	
NP-SBJ-1→NNP NNP	VP→VBD NP	13	
NP-PRD→NP PP	NP-PRD→NP SBAR	7	
NP-LOC→NNP	S-TPC-1→NP-SBJ VP	6	
NP-SBJ→NP , NP-LOC ,	NP-LOC→NP , NP	3	
NP-LOC→NNP	NP-LOC→NP , NP	3	
FRAG→NP-SBJ NP	NP-LOC→NP , NP	3	

Table 4.2: The left column has the production pairs that we identified as occurring in adjacent sentences significantly more than chance. The top 10 productions that Cheung and Penn (2010) found as repeated very often are in the rightmost column.

" Markey said we could [have done this in public " because so little sensitive information was disclosed]$_{\text{VP}}$, the aide said. [Mr. Phelan]$_{\text{NP-SBJ}}$ then responded that he would have been happy just writing a report to the panel, the aide added.

Similarly, in the adjacent sentence pairs from our corpus containing the items (NP-LOC $\rightarrow$ NNP | S-TPC-1 $\rightarrow$ NP-SBJ VP), p_1 often introduced a location name and was associated with the title of a person or organization. The next sentence has a quote from that person, where the quotation forms the topicalized clause in p_2. Here the intentional structure is INTRODUCE X / STATEMENT BY X such as in the following example:

Two years ago, the Rev. Jeremy Hummerstone, vicar of Great Torrington, [Devon]$_{\text{NP-LOC}}$, got so fed up with ringers who didn't attend service he sacked the entire band; the ringers promptly set up a picket line in protest. ["They were a self-perpetuating club that treated the tower as sort of a separate premises]$_{\text{S-TPC-1}}$, " the Vicar Hummerstone says.

These results show the existence of reasonable patterns for a domain in the syntax of adjacent sentences. Even though the Penn Treebank contains function tags and traces which are not provided by automatic parsers, we can expect that other such syntactic patterns would be present in most domains and genres. Our metric for organization quality aims to characterize syntactic patterns on a broad scale. The model relies on two assumptions which summarize our intuitions about syntax and intentional structure:

1. Sentences with similar syntax are likely to have the same intention or communicative goal.

2. Regularities in intentional structure of articles will be manifested in syntactic regularities between adjacent sentences.

Below we describe the models we developed to learn such syntactic patterns.

4.2 Representing syntax

We use two methods to represent syntax, both derived from the constituency parse of a sentence. These representations use syntax exclusively. All terminals (words) are removed from the parse tree before any processing is done. The leaf nodes in our parse trees are part of speech tags.

4.2.1 Productions

In this representation we view each sentence as the set of grammatical productions, LHS $\rightarrow$ RHS, which appear in the parse of the sentence. Since we have removed the terminal nodes from the tree, the right-hand side (RHS) contains only non-terminal nodes.

This representation is straightforward, however, some productions can be rather specific with long right hand sides. Another apparent limitation of this representation is that it contains sequence information only about nodes that belong to the same constituent.

4.2.2 *d*-sequence

We expected that the sentence (surface) order of syntactic items should be beneficial for learning patterns. For example, the fact that a sentence starts with a topicalized prepositional phrase as in our example in Table 4.1 can be captured by having sequence information about the syntactic items in the sentence. The simplest approach to satisfy this need is to represent the sentence as the sequence of part of speech (POS) tags. But this representation loses all the abstraction provided by higher level nodes in tree. Instead, we introduce a more general approach, *d-sequence* where the level of abstraction is controlled using a parameter *d*.

The parse tree is truncated to depth at most *d*, and the leaves of the resulting tree listed left to right form the *d*-sequence representation. For example, in Figure 4.1, the line depicts the cutoff at depth 2.

Next the representation is further augmented; all *phrasal* nodes in the *d*-sequence are annotated (concatenated) with the left-most leaf that they dominate in the full non-lexicalized parse tree. This annotation is shown as suffixes on the S, NP and VP nodes in the figure. The resulting representation conveys richer information about the structure of the subtree below nodes in the *d*-sequence. For example, "the chairs", "his chairs", "comfortable chairs" will be represented as NP_{DT}, $NP_{PRP\$}$ and NP_{JJ}. Note that this augmentation is different from the popular node annotation methods used for parsing models, where the head word of the phrase is used for annotation [17, 23]. We use the leftmost leaf so that we get more information about the phrase type. For example, during head annotation, all the above example phrases will obtain "chairs" or its part of speech NNS

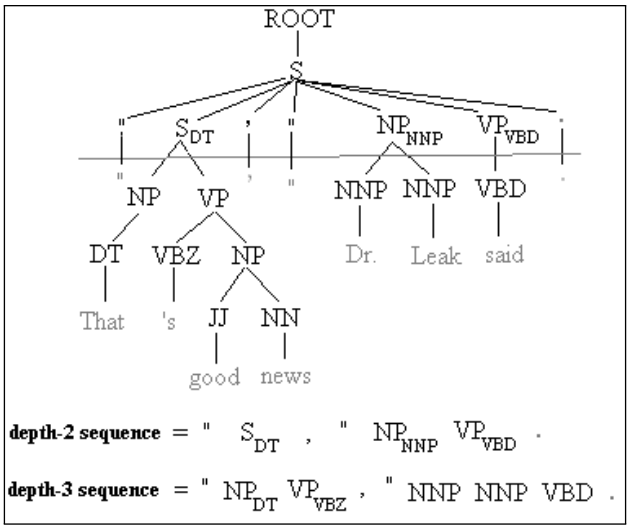

Figure 4.1: Example for *d*-sequence representation

as the label. In contrast, the left most leaf distinguishes the same phrases as definite noun phrase, having a possessive role, and descriptive noun phrase respectively. We expected that this distinction is more suitable for our task of indicating the communicative goal of a sentence.

In the *d*-sequence approach, sentences are viewed as sequences of *syntactic words* (w_1,w_2...,w_k), $k \leq p$, where p is the length of the full POS sequence and each w_i is either a POS tag or a phrasal node+POS tag combination.

Figure 4.1 shows the *d*-sequence representation for the example sentence [*"That's good news," Dr. Leak said.*]. At depth-2 (cutoff shown by the horizontal line), the representation is (w_1=" , w_2=S_{DT} , w_3=, , w_4=" , w_5=NP_{NNP} , w_6=VP_{VBD} , w_7=.) where the actual content of the quote is omitted. Sentences that contain attributions are likely to appear more similar to each other when compared using this representation in contrast to representations derived from word or POS sequence. The depth-3 sequence is also indicated in the figure.

The main verb of a sentence is central to its structure, so the parameter d is always set to be greater than that of the main verb and is tuned to optimize performance for predicting organization quality. In fact, we tune for the increment value that should be added to the depth of the main verb i.e. $d =$ depth of main verb $+$ increment. The increment value is a constant however, depending on the depth of the main verb, different

sentences would be truncated at different depths.

4.3 Predicting organization quality using syntactic regularities

Given a training set of articles with the same purpose, we use two models of coherence to learn syntactic regularities.

4.3.1 Simple co-occurrence model

In this approach, we estimate the probabilities of pairs of syntactic items from adjacent sentences in the training data and use these probabilities to compute the organization quality of new texts.

The coherence of a text T containing n sentences $(S_1...S_n)$ is computed as:

$$P(T) = \prod_{i=2}^{n} \prod_{j=1}^{|S_i|} \frac{1}{|S_{i-1}|} \sum_{k=1}^{|S_{i-1}|} p(S_i^j | S_{i-1}^k)$$

where S_x^y indicates the y^{th} item of S_x. Items are either productions or syntactic word unigrams depending on the representation. Suppose that $S_i^j = w_q$ and $S_{i-1}^k = w_r$ where w_q and w_r are syntactic items in the vocabulary. The conditional probability for the equation above is computed as follows and uses Lidstone smoothing.

$$p(w_q | w_r) = \frac{c(w_r, w_q) + \delta_C}{c(w_r) + \delta_C * |V|}$$

where $c(w_r, w_q)$ is the number of adjacent sentence pairs where the first sentence contains the item w_r and is immediately followed by a sentence that contains w_q. $c(w_r)$ is the number of sentences which contain w_r. $|V|$ is the vocabulary size for syntactic items.

4.3.2 Hidden Markov Model approach

This approach uses a Hidden Markov Model (HMM) which has been a popular implementation for modeling coherence [8, 41, 51]. The hidden states in our model depict communicative goals by encoding a probability distribution over syntactic items. This distribution gives higher weight to syntactic items that are more likely for that communicative goal. Transitions between states record the common patterns in intentional

Cluster a	Cluster b
ADJP → JJ PP | VP → VBZ ADJP	VP → VB VP | VP → MD VP
[1] This method VP-[is ADJP-[capable of sequence-specific detection of DNA with high accuracy]-ADJP]-VP .	[1] Our results for the difference in reactivity VP-[can VP-[be linked to experimental observations]-VP]-VP .
[2] The same VP-[is ADJP-[true for synthetic polyamines such as polyallylamine]-ADJP]-VP .	[2] These phenomena taken together VP-[can VP-[be considered as the signature of the gelation process]-VP]-VP .

Table 4.3: Example syntactic similarity clusters using productions representation. The top two descriptive productions for each cluster are also listed.

structure for the domain. This approach can be expected to have some benefits compared to the simple co-occurrence model. We can model document beginning and end in a better manner using the HMM and also implement more directly the idea that sentences with similar syntax could have the same intentional structure.

Parameter initialization.

In this syntax-HMM, states h_k are created by clustering the sentences from the training documents by *syntactic similarity*. For the productions representation of syntax, the features for clustering are the number of times a given production appeared in the parse of the sentence. For the d-sequence approach, the features are n-grams of size one to four of syntactic words from the sequence. Clustering was done by optimizing for average cosine similarity and was implemented using the CLUTO toolkit [174]. C clusters are formed and taken as the states of the model. Table 4.3 shows sentences from two clusters formed on the abstracts of chemistry journal articles (taken from [84]) using the productions representation. Cluster (a), appears to capture descriptive sentences and cluster (b) involves mostly speculation type sentences.

The emission probabilities for each state are modeled as a (syntactic) language model derived from the sentences in it. For productions representation, this is the unigram distribution of productions from the sentences in h_k. For d-sequences, the distribution is

computed for bigrams of syntactic words. These language models use Lidstone smoothing with constant δ_E. The probability for a sentence S_l to be generated from state h_k, $p_E(S_l|h_k)$ is computed using these syntactic language models.

The transition probability p_M from a state h_i to state h_j is computed as:

$$p_M(h_j|h_i) = \frac{d(h_i,h_j) + \delta_M}{d(h_i) + \delta_M * C}$$

where $d(h_i)$ is the number of documents whose sentences appear in h_i and $d(h_i,h_j)$ is the number of documents which have a sentence in h_i which is immediately followed by a sentence in h_j. In addition to the C states, we add one initial h_S and one final h_F state to capture document beginning and end. Transitions from h_S to any state h_k records how likely it is for h_k to be the starting state for documents of that domain. δ_M is a smoothing constant.

The likelihood of a text with n sentences is given by:

$$P(T) = \sum_{h_1 \ldots h_n} \prod_{t=1}^{n} p_M(h_t|h_{t-1}) p_E(S_t|h_t)$$

Re-estimation.

With these settings as an initial HMM, we use the Baum Welch algorithm [133] to iteratively re-estimate parameters. We run iterations until the training data likelihood no longer increases or a fixed number of iterations is reached.

All model parameters—the number of clusters C, smoothing constants δ_C, δ_E, δ_M and d for d-sequences—are tuned to optimize how well the model can distinguish well-organized articles from incoherent ones. We describe these settings in the next section.

We used the models we developed to perform text quality prediction for both of our genres related to research writing—academic articles and science journalism.

4.4 Text quality assessment for academic articles

We model the structure of only the non-experimental sections such as abstract, introduction and related work. We use data from two corpora both containing computational

linguistics articles published in ACL (Association for Computational Linguistics) conferences.

ACL Anthology Network (AAN) Corpus: [134] provides the full text of publications from ACL venues. The AAN corpus is produced through OCR analysis and the different sections of the articles are not easily identifiable. So we find the boundaries of sections using the ParsCit tagger[14] developed by Councill, Giles and Kan (2008) [28]. This tool can recover the logical structure of academic articles and also mark headers, footnotes, equations, etc. We remove the extraneous content such as footnotes, table and figure headers, equations, and examples, and keep only the main text of the articles.

We use articles from years 1999 to 2011 of ACL for creating our datasets. For training, we randomly choose 70 articles from ACL and NAACL (North American Chapter of the Association for Computational Linguistics) main conference proceedings. Similarly, we obtain a development corpus of 36 articles and a test set of 500 articles, also from ACL and NAACL conferences. We only choose articles in which all three sections—abstract, introduction and related work—could be successfully identified using Parscit.[15] This data was sentence-segmented using MxTerminator [141] and parsed with the Stanford Parser [75]. Since these articles form the main dataset which we use to demonstrate the performance of our models and for comparison with related work, we will refer to this dataset as EXPT CORPUS.

Argumentative Zoning (AZ) Corpus: This corpus was developed by Teufel (2000) [158] and has 80 ACL articles for which intentional structure was been manually annotated. These articles are taken from years 1994 to 1996 of the ACL conference and do not overlap with our training and test articles in the EXPT CORPUS above. Each sentence was assigned to one of seven *argumentative zones*. These zones are briefly defined below (text taken from Teufel, Carletta and Moens (1999) [160]).

> BACKGROUND: Sentences describing some (generally accepted) background knowledge

[14] http://aye.comp.nus.edu.sg/parsCit/

[15] We also exclude introduction and related work sections longer than 50 sentences and those shorter than 4 sentences since they often have inaccurate section boundaries.

AIM: Sentences best portraying the particular (main) research goal of the article

CONTRAST: Sentences contrasting own work to other work; sentences pointing out weaknesses in other research; sentences stating that the research task of the current paper has never been done before; direct comparisons.

BASIS: Statements that the own work uses some other work as its basis or starting point, or gets support from this other work

TEXTUAL: Explicit statements about the textual section of the paper

OTHER: Sentences describing aspects of some specific other research in a neutral way (excluding Contrastive or Basis statements)

OWN: Sentences describing any aspect of the own work presented in this paper—except what is covered by Aim or Textual

An 'undefined' label was assigned to sentences which could not be placed in any of the above categories. In the AZ corpus, sentence segmentation was already performed during annotations. We create parse trees for these sentences using the Stanford Parser.

This corpus gives us a way to examine how our model's predictions compare to actual annotations of intentional structure. Another attractive aspect of our test setup is that our EXPT CORPUS for training and testing and this AZ corpus contain articles from the same conferences. Therefore these corpora can be compared and experimented with without concerns about differences that could arise in subgenres of academic writing. Further recently, Teufel and Kan (2009) [161] have released a supervised classifier to perform zone annotations. This classifier is trained on the AZ corpus and annotates a sentence into one of the seven zones listed above. Therefore we are also able to run the classifier on our training and test sets from EXPT CORPUS and create a model for organization quality based on these predicted zones.

As described in Chapter 2.4, for the academic genre, we create approximate examples of well-organized and incoherent articles. We use pairs of articles, where one has the original document order and the other is a random permutation of the sentences from the same document. Since the original article is more coherent than a random permutation,

Section	Test pairs	Coocc prod	Coocc d-seq	HMM prod	HMM d-seq
Abstract	8815	44.0	47.2	56.8	62.9
Intro	9966	54.5	53.0	72.1	68.8
Rel. wk.	10,000	54.6	54.4	68.0	72.7

Table 4.4: Accuracy for differentiating original from permuted sections on ACL articles

we evaluate a model using the accuracy with which it can identify the original article in the pair, i.e. it assigns higher probability to the original article.

4.4.1 Accuracy of the syntax models

There are four types of syntactic models in our work: the two simple co-occurrence models using the production and d-sequence representations and the HMM models with the two representations. We train each model for each corpus and each section. We use only the EXPT CORPUS for this set of experiments. The models are tuned on the respective development data, on the task of differentiating the original from a permuted section.

The average length of abstract sections in our data is 5 sentences, introductions have 22 and related work have 21 sentences on average. To create the development corpus, we computed a maximum of 30 permutations per article and paired them with the original articles. For the test set, we created a maximum of 20 permutations for each example section to use as the negative examples.

The baseline accuracy for differentiating an original section from a paired permutation is 50%.

The number of test pairs for each section and the accuracies of the syntax models are presented in Table 4.4.

The simple co-occurrence models are rather weak. They give close to random baseline accuracies. For abstracts, the accuracies are around 44 to 47% and for introduction and related work sections, the numbers are around 54%. For this approach, there is no difference in performance depending on the method used to represent syntax. Both production and d-sequence based models have similar accuracies.

The HMM-based models give much higher accuracies than the simple co-occurrence

Dataset	No. states		d for d-seq	
	HMM-prod	HMM-d-seq	Co-occ	HMM
ACL abstracts	10	18	MVB + 3	MVB + 6
ACL intro	7	12	MVB + 6	MVB + 8
ACL relwk	11	13	MVB + 2	MVB + 1

Table 4.5: Best parameter settings for number of HMM states and d-sequence depth cutoff. MVB stands for 'depth of main verb in the sentence'

ones. The productions-based HMM model has 57% accuracy on abstracts and much higher, 72% for introductions and 68% for related work. The HMM d-seq model is overall even better performing, with accuracies consistently above 62% for all sections. Its performance on introduction sections is lower than that provided by productions representation, however, the difference between them is close, 3%. Since these tests are performed over a large number of article pairs, the improvements provided by the HMM models are significant and useful.

Among the three sections, we find that the models have better accuracies on the introduction and related work sections compared to abstracts. This result indicates that more regular patterns or easily identifiable patterns are present in introductions and related work section while the structure of abstracts is more diverse. We revisit this finding in the next section when we analyze a model for organization that we built from oracle annotations of intentional structure.

It should also be noted that our corpus is significantly more challenging compared to articles used in prior work for predicting organization quality. Our articles are longer and the ACL corpus also has OCR errors which affect sentence segmentation and parsing accuracies. In our paper on this work [95], we also report results on shorter news articles where we show that accuracies as high as 90% can be obtained using our syntax models.

The best parameter settings for our models are given in Table 4.5.

The number of HMM states is less than 20 in all cases. The depth parameter for creating the d sequences is also interesting to analyze. The maximum depth of sentences in our corpus is around 11 and on average the main verb of the sentences is at a depth of

3. Our parameter settings show that in some cases, the tree is truncated at a lower depth compared to others. Particularly, for related work sections, both simple co-occurrence and HMM models have a low value for d. The highest values for d are chosen in the case of introduction sections and they range from MVB (depth of main verb) + 6 to MVB + 8.

Overall, for the simple co-occurrence models, the trees are truncated at a lower depth. In contrast, for the HMM approach, for two out of the three sections, the depth is much greater. Therefore the nodes in the d-sequences created for the HMM models are closer to part of speech tags. One reason for this difference could be the nature in which co-occurring syntactic items are computed in each approach. In the simple model, we compute conditional probabilities for pairs of syntactic items. When each item is abstract, the information encapsulated in the pair is more general and applicable to a larger number of item pairs during testing. Hence lower depth settings and the abstract nodes provided in those d-sequences could be preferred. The HMM model on the other hand uses a language model of syntactic items within each state and fine-grained events could be more helpful for keeping the language models of the different states unique and distinct from each other.

4.4.2 Comparison with other models for organization quality

In this section, we present comparisons of our syntax models with a few other approaches. We focus on two main directions for this comparison.

Intentional structure-based

In this category, we implement two methods which are directly based on the idea of intentional structure. For building these models, we use two resources—the manual annotations available in the AZ corpus and the supervised classifier for argumentative zones developed by Teufel and Kan (2009) [161]. This classifier uses features based on lexical items, discourse connectives and some syntax information such as verb tense and part of speech tags. The classifier has an Fscore of 40% reported in that work for the seven-way classification.

a) IS-Oracle. In this model we utilize the manual AZ annotations only. We extract the

abstracts, introduction and related work sections in the 80 articles in the AZ corpus. There are 80 abstract, 65 introduction and 17 related work sections which were obtained. The counts are lower for introduction and related work sections since they were not present (with an easily identifiable heading) in all the articles. We build models for organization quality one for each section by recording the likely sequences of zones using a Markov Chain.

We follow a leave one out procedure for performing our tests. We train a Markov Chain on all but one of the examples. The transition probabilities are smoothed with Laplace method. For the held out section, we create a maximum of 20 permutations and pair them with the original section. For each sentence in these test articles, we again know the exact zone from the annotations. We use the trained model to compute the likelihood of each article in the pair and examine if the model assigns higher probability to the original ordering of zones compared to a permuted one.

Note that the training and evaluation of this model is done on only using the articles from the AZ corpus and do not involve the the training and test sets from the EXPT CORPUS. As a result, the number of test pairs and training data is different from previous section. However, this model is only built to understand the performance of an oracle method. For the other methods explained below, only the EXPT corpus is used and the results are directly comparable with our syntax models.

b) IS-Supervised. Here we use the supervized zone classifier for building a model. For each sentence in the training and test sets in the EXPT CORPUS, we use the classifier to obtain a prediction for the zone of the sentence. Then we train a Markov Chain on the predicted zones of the EXPT training set. We run this model on the permutations on the EXPT test set and compute the accuracy with which the model differentiates original and permuted sections. We smooth the transition probabilities using Lidstone smoothing and the smoothing parameter tuned on the EXPT CORPUS development data.

These two models use direct information about intentional structure compared to our methods which are based on a simple assumption about syntactic patterns. The accuracies of these methods will provide us with an understanding of the extent to which intentional structure is useful for coherence prediction.

Other models for organization quality.

Here we compare with other methods proposed in prior work to predict organization quality and which are not directly based on intentional structure or using only syntactic patterns. We choose three methods for comparison such that they are based on different aspects of writing—flow of subtopics, entity structure and reference forms.

a) Content models (CM). introduced by Barzilay and Lee (2004) [8] and Fung and Ngai (2006) [51] use lexically driven HMMs to capture organization of texts. The hidden states represent the topics of the domain and encode a probability distribution over words. Transitions between states record the probable succession of topics. Clusters are created using word bigram features after replacing numbers and proper names with tags NUM and PROP. The emissions are given by a bigram language model on words from the clustered sentences.

We implement the content models approach using our HMM implementation. After initializing the parameters in the same way as described in Section 4.3.2, we run Baum Welch iterations for a set number of times or until convergence. This re-estimation process is the main difference between our implementation and that used in Barzilay and Lee (2004) [8]. Barzilay and Lee use a Viterbi re-estimation method. They start with an initial clustering of sentences as the states. Then they obtain the best state sequence for each training article under the initial model. When the current clustering for a sentence is different than the one assigned as likely by the model, the clustering is adjusted and the sentence put in the predicted 'most likely' cluster. The training is then done with the new clustering. This Viterbi re-estimation process is repeated mutliple times before the clusters are finalized. We tune the parameters for number of clusters and smoothing parameters for this model in a similar manner as our syntax model using the development data.

b) Entity grid (Egrid). introduced in Lapata and Barzilay (2005) [80] and Barzilay and Lapata (2008) [7] is another popular approach for predicting organization quality. We described this method in detail in the related work section of Chapter 2. We use the generative model described in Lapata and Barzilay (2005) for our comparison. In this method, the entity grid is computed similar to the discriminative approach described in Chapter 2. The text is converted into a matrix, where rows correspond to sentences, in

the order in which they appear in the article. Columns are created one for each entity appearing in the text. Each cell (*i*,*j*) is filled with the grammatical role $r_{i,j}$ of the entity j in sentence i. We computed the entity grids using the Brown Coherence Toolkit[16]. Rather than use the proportion of transitions as features, in the generative approach, the probability of the text (T) is obtained as follows:

$$P(T) = \prod_{j=1}^{m} \prod_{i=1}^{n} p(r_{i,j} | r_{i-1,j} ... r_{i-h,j})$$

for m entities and n sentences. Parameter h controls the history size for transitions and is tuned during development. When $h = 1$, for example, only the grammatical role for the entity in the previous sentence is considered and earlier roles are ignored.

We also use a modified entity grid model proposed by Elsner and Charniak (2011) [44]. The model allows the use of longer histories for the entities without sparsity issues. Rather than compute conditional probabilities for a new role given the history of previous roles, this model uses logistic regression based on history features to get the probability of a new role filling the current position. We use the implementation provided with the Brown Coherence Toolkit to obtain the likelihood of articles under this model. We will refer to this model as Egrid-LogR.

c) Entity reference form (Ref). The third model we compare with under this category was introduced by Elsner and Charniak (2008) [43]. This model is based on the idea that references to discourse new entities have a clearly distinguishable form compared to mentions of entities that are already introduced in the discourse.

The implementation of this method involves two steps. Firstly, Elsner and Charniak use features developed by Uryupina (2003) [165] to predict a probability for the discourse newness of each entity mention in test article. These features are based on properties of the reference form of the entity mention. In a separate step, they identify all coreferring chains of entities in the article. The coreference is computed in an approximate manner—using head word match. These coreference chains are used to give approximate labels to entities as 'discourse-new' or 'discourse-old'. Suppose that this label is L_{np}. Then the discourse newness model developed in the first step is applied to the entities and the

[16] `http://www.cs.brown.edu/~melsner/manual.html`

Approach	Abstract	Intro.	Rel. work
Intentional-structure based			
IS-oracle (different test set)	82.9	98.9	88.2
IS-supervised	62.4	68.7	58.7
Syntax models			
HMM production	56.8	72.1	68.0
HMM *d*-sequence	62.9	68.8	72.7
Other organization models			
Content Models	68.1	76.5	54.2
Entity Grid	45.0	78.8	70.6
Entity Grid - LogR	50.7	79.1	48.9
Ref. form	56.3	72.1	63.9

Table 4.6: Accuracies of alternative methods to predict organization quality on academic articles

probability of the text T is computed as:

$$P(T) = \prod_{np:NPs} p(L_{np}|np)$$

where NPs the set of entities in the text.

Accuracies on text quality prediction

The accuracies of these different approaches together with our syntax models are reported in Table 4.6.

The oracle models work extremely well for the task. The accuracies are above 80% and even 99% for Introductions. Our claim that intentional structure patterns could be good predictors of text quality for the academic genre appears to be quite strong.

These high accuracies indicate that there are clear patterns in the intentional structure of all three sections. In Table 4.7, we show the Markov chain probabilities (computed on the training data) for these sections listing only those that are above 0.25. For this

Abstracts

	bkg	own	bas	aim	ctr	oth	txt	END
START	-	-	-	0.55	-	-	-	-
bkg	0.37	-	-	0.27	-	-	-	-
own	-	-	-	-	-	-	-	0.37
bas	-	0.57	-	0.29	-	-	-	-
aim	-	0.59	-	-	-	-	-	-
ctr	-	0.25	-	0.43	-	-	-	-
oth	-	-	-	0.42	-	0.33	-	-
txt	-	-	-	-	-	-	-	-

Introductions

	bkg	own	bas	aim	ctr	oth	txt	END
START	0.71	-	-	-	-	-	-	-
bkg	0.77	-	-	-	-	-	-	-
own	-	0.68	-	-	-	-	-	-
bas	-	0.36	-	-	-	-	-	-
aim	-	0.43	-	-	-	-	-	-
ctr	-	-	-	-	0.45	-	-	-
oth	-	-	-	-	-	0.71	-	-
txt	-	-	-	-	-	-	0.55	0.29

Table 4.7: Markov chains showing some of the top probabilities for zone transitions in academic articles

example, we did not apply smoothing while calculating the probabilities.

We see that many probabilities in the tables are quite high, the highest value in the matrix for abstract sections is 0.59 and a value even higher, 0.77, is present for the introductions. Because of these strong regularities, the oracle model makes highly accurate predictions. Particularly, in introductions, notice that the first sentence is a 'background' zone 71% of the time. Therefore during testing on permutation-based examples, a different zone for the first sentence can be easily penalized by this model leading to good prediction accuracies. Also overall, the patterns are stronger for introduction and related

work sections (in terms of probability values) and this could explain the better performance on these sections compared to abstracts. Recall that in our evaluation of the syntax models (Section 4.4.1), we had observed a similar trend with the methods obtaining better accuracies on introduction and related work sections than on abstracts.

The oracle model accuracies however came from exact knowledge about the zones for each sentence.

Our second model *IS-supervised* implements the same approach as the oracle, but based on automatic predictions of the argumentative zones. We find that the accuracies of this model are much lower, around 59% to 69%. Hence very accurate identification of zones is important for good performance. In fact, we find that IS-supervised has lower accuracies compared to unsupervised approaches such as our syntax models. The syntax models achieve 4 to 10% better accuracies.

Other organization quality methods also provide good accuracies. Particularly, the entity grid method provides the best accuracies for predicting the coherence of introduction and related work sections. The accuracy of this method on introductions is close to 80%. Content models have the best accuracy on the abstract section 68%, the second best accuracy on abstracts is obtained by the *d*-sequence HMM. The version of the Entity Grid based on logistic regression and the organization model based on reference form have very good accuracies on introduction sections and not as high on the others.

Combination of methods

The entity grid, content model, reference form model and our syntax-based approaches are based on different aspects of organization. In this section, we present experiments where we combined their predictions and examined if better accuracies result from any combinations compared to individual models. We did not use the Egrid-LogR model since it is based on a similar idea as the entity grid and did not provide any special benefits on our data. Among our syntax models, we choose the *d*-sequence HMM, since it gave the best results overall.

Combination of these models can be performed in many ways. In prior work, content and entity grid methods have been combined generatively [41] and using discriminative

training with different objectives [149]. In both settings, the combinations were found to have improved prediction accuracies. In this work, we followed a simple approach: we combine the predicted text probabilities from the models in a supervised classification system.

We did not have separate training data for combining the models. So we perform the following classification experiment which combines the predictions made by different models on the *test set*. Each test pair (article and permutation) forms one example and is given a class value of 0 or 1 depending on whether the first article in the pair is the original one or the second one. The example is represented as an n-dimensional vector, where n is the number of models we wish to combine. For instance, to combine content models and entity grid, two features are created: one of these records the difference in log probabilities for the two articles from the content model, the other feature indicates the difference in probabilities from the entity grid.

A logistic regression classifier is trained to predict the class using these features. The test pairs are created such that an equal number of examples have class 0 and 1, so the baseline accuracy is 50%. We run this experiment using 10-fold cross validation on the test set after first obtaining the log probabilities from individual models. In each fold, the training is done using the pairs from all but 10 articles and tested on permutations from the remaining 10 articles. These accuracies are reported in Table 4.8. When the accuracy of a combination is better than that using any of its smaller subsets, the value is bolded.

First, we examine combinations of models introduced in prior work. Content and entity grid methods have been reported to have complementary strengths [41, 149]. In our data, we see the improvement only for abstract sections. However, the techniques in prior work used much more specialized training to combine the methods and we can expect that greater improvements can be obtained using improved ways of combining the models. Elsner and Charniak (2008) [43] also report that entity grid and reference form models can be combined to obtained improved results. Again we find the improvement is on the abstract sections only. Content models and reference form capture clearly different aspects of writing and the improvements by combining them are present for all three sections. Finally in the case of introductions, combinations of all three models—content,

Model	Abstract	Intro	Relwk
Combination of prior models			
Content models + entity grid	**74.9**	74.8	66.8
Content models + Ref. form	**75.2**	**82.2**	**65.3**
Entity grid + Ref. form	**67.9**	71.7	68.9
Content models + entity grid + Ref. form	74.7	**83.6**	63.9
Combination of prior models with syntax HMMs			
Content models + HMM-*d*-seq	**71.6**	**79.8**	71.6
Entity Grid + HMM-*d*-seq	**65.0**	71.7	**78.2**
Ref. form + HMM-*d*-seq	**70.2**	**79.5**	**76.8**
Content models + entity grid + Ref. form + HMM-*d*-seq	75.3	**86.3**	75.8

Table 4.8: Performance of previously proposed methods to predict organization quality and the results when they are combined with the syntax-based models

entity grid and reference form—is better than combinations of any two models. This result was not obtained for the abstract and related work sections.

The syntax HMM model is complementary with reference form for all three sections. This result indicates that coreference and discourse information can improve our syntax method based mostly on the structure of sentences. When combined with entity grid, the accuracies improve over the individual models for abstract and related work. With content models the improvements are for abstracts and introductions. Therefore the syntax HMM has new information compared to all these prior models. However, combinations of the syntax HMM with all the three other models showed better results than combinations of two models only for introductions.

While we followed a simple approach for combining the methods, we expect that other ways to understand the differences between these models and how to combine them effectively is a good direction for future work.

4.4.3 Syntax-based models and intentional structure

So far we assumed that syntax provides a rough proxy for intentional structure and used this idea as a motivation for developing models based on syntax. In this section, we use the zone annotations from the AZ corpus to test how far this assumption works and if sentences with similar syntax show indications of the same communicative goal.

We compare the predictions of three models with the zone annotations—the HMM-based syntax models using the production and d-sequence representations and the HMM-based content models. The syntax and content models differ on the basis by which organization quality is captured but have a similar implementation. We expected that by comparing these two types of methods with the zone annotations, we can understand if syntax is more indicative of intentional structure categories compared to lexical patterns.

We use the tuned HMM models of each type which we created for the experiments in the previous section. We only use the HMM models since for these, we are able to obtain the likely state sequence for each article and compare these state labels on the sentences with the manual zones labels provided for them. For the co-occurrence-based syntax models, it is not straightforward to obtain any labels for individual sentences. Further their accuracies are much lower and hence of little interest for this analysis.

We setup the comparison as follows. In the AZ corpus, we separate out the abstracts, introduction and related work sections. There are 80, 65 and 17 sections of each kind respectively. From each HMM model (trained on the EXPT CORPUS), we obtain the best state sequence for each section using Viterbi decoding. We use the state for each sentence as a label under the HMM model. Similarly, we have a set of gold labels for each sentence which are the zones that were annotated in AZ.

The distribution of gold standard labels is shown in Table 4.9. There is a highly skewed distribution for abstracts, with 46% of sentences belonging to the 'own' zone. Similarly, 51% of sentences in related work sections are in the 'other' zone.

The number of states in each of the HMM models is reported in Table 4.10. Each state is a possible label for the sentences. The content models have more states than the syntax ones.

We treat each set of labels (state or zone) as the output of a clustering method and use

Section	No. sents	aim	basis	backg.	contrast	other	own	textual
Abstract	356	101	7	30	28	24	166	0
		(28.4)	(2.0)	(8.4)	(7.9)	(6.7)	(46.6)	(0.0)
Introduction	1417	97	42	363	163	338	332	82
		(6.8)	(3.0)	(25.6)	(11.5)	(23.9)	(23.4)	(5.8)
Related work	444	11	13	59	72	227	52	10
		(2.5)	(2.9)	(13.3)	(16.2)	(51.1)	(11.7)	(2.3)

Table 4.9: The number and percentage (in parentheses below) of sentences in different zones in the AZ corpus. The seven zones are described in the beginning of Section 4.4. The total number of sentences in the texts for a section are under 'no. sents' column.

Section	HMM production	HMM d-sequence	Content model
Abstract	10	18	40
Introduction	7	12	17
Related work	11	13	23

Table 4.10: The number of states in syntax models and content model

cluster comparison metrics to measure the similarity of the two labelings.

We use three metrics for this analysis–Adjusted Rand Index (ARI) [65, 138], Jaccard Coefficient (JC) and cluster purity (CP).

The first two measures are based on pairs of items and examine how these pairs are placed in the two given clusterings. Let us call the clusterings as C_A and C_B. There are four possible placements for any pair of items:

SS - the pair belong to the same cluster in both C_A and C_B

DD - the items in the pair belong to different clusters in both C_A and C_B

SD - the pair belong to the same cluster in C_A but different ones in C_B

DS - the pair belong to the same cluster in C_B but different ones in C_A

Let us refer to the count of pairs falling in each of these settings as c(SS), c(DD) and so on.

A simple metric for similarity of clusterings can be computed using the Rand Index [138] which is defined as:

$$\text{Rand Index} = \frac{c(SS) + c(DD)}{c(SS) + c(DD) + c(SD) + c(DS)}$$

The Rand Index measures how many pairs have agreeing clusterings in the two methods and its value ranges from 0 (no concordances between the clustering) to 1 (same set of clusters). However the expected value for the Rand Index of two random partitions is not a constant value. So Hubert and Arabie (1985) [65] proposed a method to correct the index for agreement due to chance. In the Adjusted Rand Index the expected value for the index of two random clusterings is 0. The measure itself varies between -1 and 1, where positive values indicate agreement of clusters greater than chance and negative values indicate lesser agreement than chance.

One problem with ARI is that it gives credit for the DD pairs where the clustering correctly places items that should not be in the same clusters. But typically the number of DD pairs is large and overwhelms the other counts. The Jaccard Coefficient addresses this issue by only giving credit for placing a pair of items in the correct and same cluster. The DD counts are ignored from the equation.

Model	Adj Rand Index			Jaccard			Cluster Purity		
	Abs	**Intro**	**Rel. wk**	**Abs**	**Intro**	**Rel. wk**	**Abs**	**Intro**	**Rel. wk**
HMM-prodn	0.00	0.01	-0.02	0.10	0.14	0.15	0.48	0.29	0.51
HMM-*d*-seq	0.00	0.01	0.00	0.06	0.11	0.12	0.50	0.29	0.51
Content models	0.02	0.04	0.00	0.06	0.14	0.08	0.62	0.36	0.52

Table 4.11: Cluster metrics comparing different coherence models with argumentative zone annotations. The number of sentences in abstracts set is 356, introductions 1417 and related work 444.

$$\text{Jaccard coefficient} = \frac{c(SS)}{c(SS) + c(SD) + c(DS)}$$

Both ARI and JC easures however penalize a clustering which produces finer-level clusters of the gold standard classes. A finer level clustering reduces the value of c(SS) and increases c(SD) and the resulting ARI becomes low. So we also compute another simple measure—cluster purity. In this method, each cluster from a candidate clustering is given the label of the gold standard cluster with which it maximum overlap. Once these labels are assigned, we can compute the percentage of correct labels.

$$\text{Purity} = \frac{1}{N}\sum_{k}\max_{j}|w_k \cap c_j|$$

where $w_1, w_2 ... w_K$ are the clusters of a candidate clustering and $c_1, c_2 ... c_J$ are the gold standard clusters. The intersection gives the number of items in common between w_k and c_j. N is the total number of items.

Table 4.11 gives the metrics computed for our comparison of syntax and content models with the gold standard zone annotations

With the corrected rand index (ARI), we find that the values are all zero predicting that any concordant pairs could be simply due to chance. The Jaccard coefficient shows a different trend. The values for all models are overall low but we see that the HMM production models appear to have greater agreement with the zone annotations compared to the *d*-sequence and the lexical model. For introduction sections, the JC for HMM-productions is 0.10 and 0.06 for the other models. For related work, the productions

Model	Adj Rand Index			Jaccard			Cluster Purity		
	Abs	**Intro**	**Rel. wk**	**Abs**	**Intro**	**Rel. wk**	**Abs**	**Intro**	**Rel. wk**
HMM-prodn	0.01	0.02	0.02	0.05	0.15	0.09	0.18	0.39	0.22
HMM-*d*-seq	0.01	0.02	0.01	0.04	0.12	0.07	0.19	0.39	0.23

Table 4.12: Cluster metrics comparing the syntax models with content models. The number of sentences in abstracts set is 356, introductions 1417 and related work 444.

model has 0.15 JC, while the value is only 0.08 for content models and 0.12 for *d*-sequence method. With regard to cluster purity, the content models have better values compared to the syntax ones.

These results show that there is some evidence that syntax could indicate intentional structure but we did not obtain strong results for the comparison. Further evaluation is needed to understand how sentence types predicted by syntax or content approaches are different and how they compare with intention structure annotations. In fact, we may also expect that these approaches will not capture the same distinctions as the manual annotations. For example, the AZ scheme defines three zones, 'basis', 'contrast' and 'other', which are all related to description of previous work. The distinction between these categories is either based on whether the opinion is neutral or biased or on how the previous work is used in the paper (as starting point—'basis' zone or to validate the novelty of current work—'contrast' zone). We do not expect that syntax models can provide these distinctions. However, we hope that our models can discover useful application-based categories of sentences.

At this point, it is also interesting to look at how content and syntax models differ using the same cluster comparison measures. We treat the labels from the content model as the gold standard, and compare each of the syntax model labels against it. Table 4.12 shows these results.

These values are also not high. Therefore it appears that the content and syntax models capture rather different aspects of organization and could be a reason for their complementary nature which we observed in Section 4.4.2.

4.5 Text quality assessment for science journalism articles

Science news articles also describe research and therefore the idea of intentions is also applicable to this genre. However, the structure of science news articles is not as rigid as conference publications and a direct annotation based approach for intentional structure would be rather difficult to employ. Therefore our approximate approach is particularly valuable for this genre.

Further, while our experiments on academic articles were based on a simple idea of using permutations for incoherent samples, here we have more realistic examples. The articles in the the TYPICAL category (from our corpus described in Chapter 2) are of lower text quality overall compared to the articles in the GREAT and VERY GOOD category. While permutations gave us easily available data for experiments, they are often easy incoherent samples. This evaluation on science news articles gives us a way to go beyond simple permutation examples for testing our models.

In this section, we examine the performance of syntax models for distinguishing the text quality categories on our corpus and also compare the accuracies with those obtainable from the content models and entity grid methods.

We use the 63 articles from the GREAT category as the training set for creating the HMM models involved in the comparison. A separate model is built for the productions and the d-sequence based syntax models. We also build a content model on the same data. The number of clusters, depth, and smoothing parameters were tuned using the development set of 1000 topically matched pairs of VERY GOOD and TYPICAL articles (see Section 3.5 for details about the development corpus). During tuning, we considered a prediction as correct when the perplexity under the model for the VERY GOOD article is lower than that for the typical one. We use perplexity rather than probability since the articles in the pair have different lengths compared to the setting of permutations where both examples are equal in length. The best parameters were used to train the final model. The entity grid model is not trained on this data since we use a discriminative approach using features from the models and 10 fold cross validation on a separate dataset. So the HMM models require a separate training set for creating the HMMs, the entity grid does not. More details are explained below.

We evaluate the model on two tasks on this corpus: differentiating text quality for articles from any topic and for articles with the same topic. We use the any-topic and same-topic datasets introduced in Chapter 3 with a slight modification. In those sets, the GREAT articles were also included in the test sets. Since we have used them for training the HMM models, we create new datasets removing the GREAT articles (and articles that were topically matched with these articles in the topic normalized corpus) from the test sets. The final test sets we use have:

Any-topic: 4090 articles of each type, VERY GOOD and TYPICAL category.

Same-topic: 40900 pairs of topically matched VERY GOOD and TYPICAL articles.

The baseline random accuracies in both cases is 50%.

Using only the probability or perplexity from the HMM models for making a decision led to low performance during development. So we designed a discriminative approach with more fine-grained features. For each article in the test sets above, we obtained the best state sequence under each HMM model using Viterbi decoding. Then we compute the proportion of sentences in the article that belong to each state. Each proportion is a feature. For the entity grid, we use the discriminative approach proposed in Barzilay and Lapata (2008) [7]. The proportions of different types of entity transitions between adjacent sentences in the article are added as features (a total of 16 features). (See section 2.2 for details about the entity grid features.)

In the 'same-topic' setup, every test example is a pair of articles. The features for a test pair here is computed as the difference in each feature value from the two constituent articles. The test pairs are created such that in half the pairs the first article is the better one and in the other half, the good article is second one in the pair.

For the 'any-topic' setup, there is no pairing and the features are directly used.

We performed 10-fold cross validation in both tasks using our syntax model features separately and combined with content and entity grid methods. The results and parameter settings are shown in Table 4.13.

The productions-based syntax model gives an accuracy of 60% for the any-topic task and 62% for same-topic. The d-sequence HMM has 2% lower accuracies on both tasks. These accuracies are lower compared to results on the academic articles. But there we

Model	**Any-topic**	**Same-topic**
HMM prod (23 states)	60.6	62.9
HMM d-seq (21 states, d = MVB + 5)	58.1	60.5
Content models (31 states)	64.3	65.6
Entity grid	61.4	58.2
Content models + Entity grid	66.0	69.8
Content + Egrid + HMM prod	68.6	72.9
Content + Egrid + HMM d-seq	68.1	73.2
Content + Egrid + HMM d-seq + HMM prod	68.5	75.0

Table 4.13: Accuracy of different organization models for text quality prediction on science news articles

used permutations based data while the VERY GOOD and TYPICAL categories in this experiment are actual examples of articles of good and average quality. So the lower accuracies are expected. Still our results are 10% above the baseline indicating that the distribution of the distribution of sentence types is a valuable indicator of text quality for science journalism articles as well.

The entity grid has accuracies similar to syntax models, around 60%. Content models are better than both syntax and entity grid, obtaining about 65% accuracy in both tasks. The combination of entity grid and content model features is beneficial as expected. The accuracy is 66% for any-topic and 69% for same-topic setup.

Next we added the syntax features to the combination of entity grid and content model features. Adding either the production HMM or the d-sequence HMM both lead to improvements in the combined model (2% increase for any-topic and 3% for same-topic). The production and d-sequence HMMs also show indications of complementary nature. In the same-topic setup, the combination of all four models—content, entity grid, production and d-sequence HMMs gives the best overall performance compared to smaller subsets of features.

Again these experiments confirm that the syntax models are useful for text quality prediction and are complementary with methods introduced in previous work.

4.6 Future work

Our syntax models were complementary to entity based and content based models for organization quality. We also showed in the previous section that the information used by content and syntax models could be quite different. Given that now we have methods from prior work based on coreference, lexical statistics and reference form, and our work uses syntax, it would be quite useful in future work to obtain more qualitative results regarding how these methods and their predictions differ. Such results can shed light on the relationship between topic segments, intentional structure and entity repetition which have been proposed as components of coherent text in theoretical work [57] but there are few computational studies looking at their interaction. Understanding the relationships between these approaches will help us to combine them with greater success.

There are also some ways in which our syntax models themselves can be improved. In the two representations we have used, productions and d-sequence, we apriori choose the granularity and features for computing syntactic similarity. An interesting next step would be to compute similarity based on tree kernels [24] which will allow us to compare many different structures in the trees for similarity computation. We expect that an improved approach for computing similarity between sentences will lead to better accuracies from our approach.

However, we could only obtain weak results about the relationship between syntax and intentional structure. This is another avenue for more research in future. Currently, supervised methods to predict intentions for sentences are mostly based on lexical patterns, hand-crafted [159] or n-gram based [59, 161], and use only little syntax information related to part of speech tags and verb tense. We can experiment adding features from our syntax models to such classifiers and examine if they improve the classifier accuracy. Such an evaluation will provide a better understanding of our hypothesis that sentences which have similar syntax could have similar communicative goals.

The evaluation of our method, and also other models for organization quality will

benefit further from the creation of datasets with text quality ratings. For the academic articles, we have used permutations-type incoherent examples. But these samples are the simplest cases of low quality writing. It would be interesting to apply our models to data where direct annotations of organization quality are available. As a simpler evaluation, we can also use other proxies as the gold standard. For example, we can analyze how the probabilities assigned by our model to an academic article relate to the number of incoming citations for the article. Such an experiment will help us to understand if the quality of writing in an article has any noticeable relationship with traditional influence metrics. If there are such indications, text quality measures will be a valuable component we may wish to add to search systems for these articles. Current approaches for retrieving academic articles are based mostly on relevance to queries and citation counts.

4.7 Conclusions

In this chapter, we described how we built a measure for organization quality inspired by the idea of intentional structure of an article. We showed how rather than predefining and annotating intentional structure for different genres of articles, which would be too challenging in practice, we can use the syntactic structure of the sentence as a rough proxy for its intention. The models for organization built using this intuition perform well on coherence prediction tasks for both academic writing and science journalism articles. We also showed that this measure can augment existing organization metrics based on other aspects of discourse such as subtopic structure, entity coherence and entity reference form.

Further, using annotated corpora for intentional structure from conference publications, we were able to study how far our assumption of syntactic similarity and intentional structure is borne out in the annotated data. We found that there are indications that syntax could be indicative however, the signals are not strong given the current annotation corpus that we used.

Chapter 5

A classifier for text specificity

Details work together to clarify and expand the main.

[Ideas and Development trait (Section 2.1)]

The entire piece has a strong sense of balance. Key ideas stand out.

[Organization trait]

All the sentences of an article do not convey information in the same manner. Sentences in the opening paragraphs are general giving an overview of the topic. The details on the topic come later on. Finally the end of the article provides some abstraction and here the content is often general. The points above taken from the definition of the *Ideas and Development* and *Organization* traits of the Six Traits model are based on this switch between overview and detailed information in the text.

Consider the sentences in Table 5.1 taken from a news article.

Sentence (a) describes the unpopular features of the books chosen for the Booker

a) The novel, a story of Scottish low-life narrated largely in Glaswegian dialect, is unlikely to prove a popular choice with booksellers who have damned all six books shortlisted for the prize as boring, elitist and—worst of all—unsaleable.
...
b) The Booker prize has, in its 26-year history, always provoked controversy.

Table 5.1: Example general/specific sentences from news

prize and also talks more specifically about one of the selected books. Sentence (b) is the last sentence of this article and summarizes the negative sentiment by mentioning that controversy surrounding the prize is also longstanding and happens almost every year. The level of detail is markedly different in the two sentences. Sentence (b) only gives the topic. If this sentence is presented by itself, it will make a reader wonder why such a statement is made by the author. In other words, sentence (b) needs some substantiation from other parts of the text. On the other hand, sentence (a) does not create such expectations. It has details and specific information on the topic. In this work, we call sentences like (a) above as specific, while sentences of second type are called as general.

It is intuitive and noticeable that texts have a mix of such general and specific information.

Studies on academic writing [155] have identified that a hourglass-like structure is present in academic articles where the introduction and conclusion present general content and the experimental sections in between contain a lot of details. Large scale annotations carried out for discourse relations also indicate that sentences have different specificity levels. For example, in the Penn Discourse Treebank (PDTB) corpus [128], the *Instantiation* and *Restatement* relations appear to be relevant to this phenomenon. The definition of these relations from the PDTB manual is given below. *Arg1* and *Arg2* refer to the two text spans that are connected by the relation.

- Instantiation: Arg1 evokes a set and Arg2 describes it in further detail. It may be a set of events, reasons or a generic set of events, behaviors and attitudes. The relation involves a function which extracts the set of events from the semantics of Arg1 and Arg2 describes one element in the extracted set.

- Restatement: The semantics of Arg2 restates that of Arg1. The subtypes "specification", "generalization", and "equivalence" further specify the ways in which Arg2 restates Arg1. In the case of specification Arg2 describes the situation in Arg1 in more detail.

These definitions indicate that one sentence may be written to be more general than another sentence. The general sentence creates the need for more specific details which

is fulfilled by the subsequent (in the case of Instantiation and Restatement-Specification relations) sentence. Some example Instantiation and Specification relations between adjacent sentences are shown in Table 5.2.

Apart from the PDTB, other discourse frameworks such as Rhetorical Structure Theory (RST) [97] and Segmented Discourse Representation Theory (SDRT) [3] also note that sentences involved in certain discourse relations have varying degrees of specificity. We discuss the specificity differences reported in the RST and SDRT theories in further detail in the related work section (Section 5.8).

Given these observed regularities in the occurrence of general and specific information in texts, we hypothesize that specificity patterns will be useful for predicting the quality of an article. The content quality rubrics in Section 2.1 point out that the presentation of details has significant influence on quality. Too much general or specific content could make an article difficult to read. Similarly, the placement of general and specific content influences the organization quality of an article. When general content is presented without particular details, the article could appear ambiguous and on the other hand, specific information without appropriate topic statements and summaries would leave the reader without a high level understanding of the article. This chapter presents a metric for content quality and organization quality based on the idea of specificity.

To this end, we develop a supervised classifier to identify general versus specific sentences and use the predictions for analysis of text quality. Our classifier is trained on sentences from news articles. Based on the specificity differences noted in the PDTB Instantiation and Specification discourse relations, we create proxy examples of general and specific sentences from these relations. We use this data as a training corpus. We also obtain manual annotations from people for the general-specific distinction and test how the classifier trained on proxy examples performs on the direct annotations for specificity. Sections 5.2 to 5.4 provide details about the corpora and classification approach. This classifier has a high accuracy of 75% for identifying general and specific sentences. We calculate a measure for specificity of a text based on the classifier's predictions (described in Section 5.5).

We apply this automatically computed specificity measure to perform text quality

Instantiations

1.*The 40-year-old Mr. Murakami is a publishing sensation in Japan.* A more recent novel, "Norwegian Wood" (every Japanese under 40 seems to be fluent in Beatles lyrics), has sold more than four million copies since Kodansha published it in 1987.

2. *Sales figures of the test-prep materials aren't known, but their reach into schools is significant.* In Arizona, California, Florida, Louisiana, Maryland, New Jersey, South Carolina and Texas, educators say they are common classroom tools.

3. *Despite recent declines in yields, investors continue to pour cash into money funds.* Assets of the 400 taxable funds grew by $1.5 billion during the last week, to $352.7 billion.

Specifications

4. *By most measures, the nation's industrial sector is now growing very slowly—if at all.* Factory payrolls fell in September.

5. *Mrs. Hills said that the U.S. is still concerned about 'disturbing developments in Turkey and continuing slow progress in Malaysia.'* She didn't elaborate, although earlier U.S. trade reports have complained of videocassette piracy in Malaysia and disregard for U.S. pharmaceutical patents in Turkey.

6. *Alan Spoon, recently named Newsweek president said Newsweek's ad rates would increase 5% in January.* A full, four-color page in Newsweek will cost $100,980.

Table 5.2: Example Instantiation and Specification relations from the PDTB. The Arg1 of each relation is shown in italics.

assessment in two genres—summarization and science journalism.

Since summaries are a condensed version of the source articles, they cannot contain all the details from the source. Some content should be made more general than how it appears in the source. Therefore text specificity could have direct relevance for the task of summarization and we expected that the degree and placement of general and specific information could have a noticeable impact on text quality in this genre.

In fact, several studies in the summarization field have noted specificity differences in summaries. Jing and McKeown (2000) [69] manually analyzed human-written summaries in combination with their source documents. They pointed out that people in fact convert some source sentences into more general content for the summaries. But the opposite transformation is also done, some sentences become more specific than the source. But it is not known how often these transformations occur and if they impact the quality of summaries. Summarization evaluation has traditionally concerned itself with assessing content quality solely on the basis of how much important information is provided by the summary. Aspects such as how the information is conveyed has received little if any focus.

But recently, Haghighi and Vanderwende [61] built a topic model based summarization system that could select content based on both a general content distribution and on distributions of content for specific subtopics. They report that using the general distribution yielded summaries with better content than using the specific topics. The approach was later improved by Mason and Charniak [101] who modified the model's objective function to directly implement the idea that general content should be preferred. Given an input set which contains multiple documents, their objective function favors content that appears across multiple input documents and penalizes content that is specific to individual documents in the input. But the relationship between content specificity and quality of the summaries has not been studied so far in a direct manner across several systems and from the point of view of how people summarize articles.

Similarly, we expect the general-specific nature of content to be relevant for research writing. As we pointed out earlier, conference articles have been observed to be structured like a hour-glass with regard to general-specific nature. While research papers are written

for an expert audience and have such a structure, we believe that text specificity could be even more relevant for analyzing science journalism articles. The audience for science news are non-experts and proper substantiation and topic statements are necessary to guide a reader through difficult concepts. Therefore we also perform evaluations of text quality for the science journalism articles using the specificity metric. We have not used specificity features for the academic writing genre since our training data has been chosen exclusively from news articles.

5.1 Defining specificity

We define the general-specific distinction in the following way. Texts have a mix of general and specific content where:

- The general content provides high level information. They are topic statements and discuss an issue without giving much details. Specific content is related to details present in a text and provides substantiation for the issues mentioned in the general content.
- A general sentence creates an expectation for specific information. In other words, when a reader encounters a general sentence, he needs other portions of the article, before or after the general sentence, to provide substantiation and evidence for the content presented in the general sentence.

This definition of text specificity is motivated by ideas for analyzing writing quality. Advice on writing [2, 155] frequently emphasize that topic statements are important in a text and also that they need to be supplemented with details. So we wanted to analyze the extent to which the proportion and interaction between general or specific information is indicative of quality differences. There are a few other related distinctions made in prior work. We discuss them in Section 5.8.

The following sections describe a classifier that we built to identify general and specific information. Using the predictions of the classifier we then created features for text quality prediction.

5.2 Data

Our classifier is designed for sentence-level text spans and aims to make a binary distinction– general or specific. There were no existing annotations for the general-specific distinction before our work. Therefore we use two sources of examples for general and specific sentences for building our classifier. One of them is collected in an approximate matter and the other is obtained by direct annotations of the distinction.

5.2.1 From discourse relations

Given the patterns we discussed above in the Instantiation and Specification discourse relations from the PDTB, we used them to create an approximate dataset for the general-specific distinction. We consider the first sentence of these relations as an example of general sentence and the second as a specific one. Although the definitions of these relations describe the specificity of one sentence relative to the other, we do not focus on this pairwise difference in specificity. We believe that the realization of a general sentence should have some unique properties regardless of the particular sentence that precedes or follows it.[17] We also validate this hypothesis in a later section (5.4.2) by asking annotators to mark sentences from these relations as general or specific.

The PDTB annotations cover 1 million words from Wall Street Journal (WSJ) articles. Instantiations and Specifications are fairly frequent (1403 and 2370 respectively). Each relation gives rise to two examples, one general and one specific sentence. The baseline accuracy for random prediction on this data is 50%.

5.2.2 From direct annotations

We chose approximately 300 sentences each from three sources of news, Wall Street Journal [100], AQUAINT Corpus [54], and New York Times science section (articles from our science journalism corpus).

[17] We use only the *implicit* relations from the PDTB; ie, the sentences are not linked by an explicit discourse connective such as 'because' or 'but' that signals the relation.

AQUAINT: We chose 8 articles from the AQUAINT corpus [54] which is traditionally used for question answering and summarization. Six of them are news reports published by Associated Press and two are from Financial Times. Most articles here are short and we enforced a minimum length limit of 30 sentences. There are 292 sentences in the 8 articles combined. [docid: AP880713-0175, FT931-3664, AP900131-0200, FT923-5589, AP901019-0072, AP891116-0035, AP890922-0007, AP881002-0048]

WSJ: The Wall Street Journal corpus [100] has mostly finance news articles. We chose three articles from the WSJ and these are longer than those from AQUAINT, each about 100 sentences. The set has a total of 294 sentences. [docid: wsj-0445, wsj-1037, wsj-1394]

NYT-science: We chose three articles reporting science news from the science news corpus introduced in Chapter 3. While still news, these articles are quite different compared to the rest. For example, one of the articles discusses how the concentration of carbon dioxide in the atmosphere has changed over time. A total of 308 sentences were annotated from this source. [docid: 2002-03-05-1373005, 2006-11-07-1802956, 2007-05-10-1846387]

The WSJ contains mostly finance news while AQUAINT focuses on general news events.

We provided the sentences to annotators on Amazon Mechanical Turk[18]. Each sentence was annotated by five different assessors. They marked a sentence as either general, specific or "cannot decide". We briefly described the difference between general and specific sentences and gave examples. The assessors largely relied on their intuition to mark the distinction. We provided the following instructions.

> "Sentences could vary in how much detail they contain. One distinction we might make is whether a sentence is general or specific. General sentences are broad statements about a topic. Specific sentences contain details and can be used to support or explain the general sentences further. In other words, general sentences create expectations in the minds of a reader who would definitely need evidence or examples from the author. Specific sentences can stand by themselves. For example, one can think of the first sentence of an article or a paragraph as a general sentence compared to one which appears in the middle. In this task, use your intuition to rate the given

[18] http://sites.google.com/site/amtworkshop2010/

sentence as general or specific.[19] Some examples are provided below but they do not cover all the sentence types you may encounter."

Examples: (These examples were taken from New York Times science section but are different from the articles given for annotation.)

General sentences:

[G1] A handful of serious attempts have been made to eliminate individual diseases from the world.
[G2] In the last decade, tremendous strides have been made in the science and technology of fibre optic cables.
[G3] Over the years interest in the economic benefits of medical tourism has been growing.

Specific sentences:

[S1] In 1909, the newly established Rockefeller Foundation launched the first global eradication campaign, an effort to end hookworm disease, in fifty-two countries.
[S2] Solid silicon compounds are already familiar–as rocks, glass, gels, bricks, and of course, medical implants.
[S3] Einstein undertook an experimental challenge that had stumped some of the most adept lab hands of all time–explaining the mechanism responsible for magnetism in iron.

Since the same annotators did not provide judgements for all the sentences, we do not compute the standard Kappa measure. Rather in Table 5.3, we present statistics on the number of sentences split by how many annotators agreed on the sentence class.

For about two-thirds of the examples ($\sim$200) in each corpus, there was either full agreement among the five annotators or one disagreement. These results are reasonable for a task where annotators relied mainly on intuition.

It is also informative to analyze the agreement numbers split by general/specific distinction. We wanted to know if agreement is higher for one of the sentence types. Table 5.4 reports the agreement per category for the three data sets. On NYT and WSJ sentences, the judges have similar agreement on examples from both general and specific

[19] An option of selecting "cannot decide" was also given to the assessors.

Agree	WSJ	AQUAINT	NYT-science
5	96	108	82
4	102	91	121
3	95	88	102
undecided	1	5	3
Total	294	292	308

Table 5.3: Annotator agreement for general-specific distinction

WSJ			AQUAINT			NYT-science		
Agree	**General**	**Specific**	**Agree**	**General**	**Specific**	**Agree**	**General**	**Specific**
5	51 (31.8)	45 (33.8)	5	33 (28.2)	75 (44.1)	5	32 (25.6)	50 (27.7)
4	57 (35.6)	45 (33.8)	4	35 (29.9)	56 (32.9)	4	48 (38.4)	73 (40.5)
3	52 (32.5)	43 (32.3)	3	49 (41.8)	39 (22.9)	3	45 (36.0)	57 (31.6)
Total	160	133	Total	117	170	Total	125	180

Table 5.4: The annotator agreement numbers split by type of majority class

class. On the AQUAINT corpus, the agreement on the general sentences is lower than that on the other sets (58% of general sentences have agreement 4 or 5) but the agreement is considerably better when the sentence is specific (77% are at levels 4 or 5). So the specific sentences from the AQUAINT corpus appear to be easier for annotators. But on the whole, our judges made reliable judgements on both general and specific sentences.

In Table 5.5, we present example sentences with full agreement and those with low agreement from our three datasets.

The sentences with lower agreement appear to exhibit a genuine mix of general and specific characteristics. For example the first specific sentence with agreement level 3 has details about the year of the event and the people involved but the event itself is not specified. Similarity the first general sentence with low agreement has detailed description of the geologist but the findings that he reports are fairly general. This evidence from the annotators indicates that the distinction between general and specific can be treated more transparently as a matter of degree rather than as fixed binary classes.

We also observe the influence of context. Since the sentences are annotated out of con-

	Agreement 5
General	[NYT] Climatologists and policy makers, they say, need to ponder such complexities rather than trying to ignore or dismiss the unexpected findings. [AQ] There are two standard explanations why a weak dollar prompts bond prices to fall. [WSJ] In the private sector, practically every major company is setting explicit goals to increase employees' exposure to computers.
Specific	[NYT] Isabella Bailey, Anya's mother, said she had no idea that children might be especially susceptible to Risperdal's side effects. [AQ] WAAY reported at least one person died when the roof of a business collapsed from winds that overturned cars in the area. [WSJ] Apple didn't introduce a kanji machine – one that handles the Chinese characters of written Japanese – until three years after entering the market.
	Agreement 3
General	[NYT] "The geologic record over the past 550 million years indicates a good" correlation," said Robert A. Berner, a Yale geologist and pioneer of paleoclimate analysis. [AQ] He accomplished the same feat in 1980 and became the first man to sweep the events twice. [WSJ] As with many other goods, the American share of Japan's PC market is far below that in the rest of the world.
Specific	[NYT] In 2004, Dr. Berner of Yale and four colleagues fired back. [AQ] East Germany had 102 medals and 37 gold, and the United States 94 medals and 36 gold. [WSJ] "If it were an open market, we would have been in in 1983 or 1984," says Eckhard Pfeiffer, who heads Compaq Computer Corp.'s European and international operations.

Table 5.5: Example general and specific sentences with agreement 5 and 3

Corpus	General	Specific
WSJ	160 (54.6)	133 (45.4)
AQUAINT	117 (40.8)	170 (59.2)
NYT-Science	125 (41.0)	180 (59.0)
Total	402 (45.4)	483 (55.6)

Table 5.6: Distribution of general and specific sentences in the annotated data

text, sometimes, the sentences can be interpreted as general because they have pronouns and other links which appear unspecified but would be easily clear given surrounding sentences. For example, in the second specific sentence with low agreement (in Table 5.5), details about which medals were won are reported but one does not know the sports event they are associated with. When this information is also presented, we can expect that annotators might rate this sentence as specific with much more agreement. In future annotations, we plan to have a dedicated class for this type of lack of specificity. Such extended distinctions would be helpful for summarization and question-answering systems which will obviously benefit from being able to identify sentences whose interpretation relies on context.

For initial classification experiments, we consider all sentences with majority annotation (at least 3 annotators out of 5 agreed on the class) 'general' as general sentences and similarly for specific sentences. The number of general and specific sentences for each of our corpora are shown in Table 5.6. For AQUAINT and NYT-science there are about 20% more specific sentences than general. WSJ has an opposite trend with more general sentences. Overall, there are 45% general and 56% specific sentences. So a baseline prediction of majority class (specific) will give an accuracy of 56% on this data.

5.3 Features

Based on a small development set of 10 examples from Instantiation and Specification sentences, we came up with several features that distinguished between the specific and general sentences in the sample. We observed that in general sentences, strong opinion or

Sentence length - counts of words and nouns **Polarity** - counts of positive, negative, polar words both normalized by sentence length and without **NE+CD** - counts of numbers and named entities **Syntax** Counts of adjectives, adverbs, adjective phrases, adverb phrases, verb phrases, prepositional phrases, average length of verb phrases	**Word specificity** - min, max, average length of path from a noun or verb to root of WordNet through hypernym relations - avg, min, max Inverse Document Frequency (IDF) of words in the sentences. IDF was computed from one year of New York Times news articles. **Language models** - Log probability and perplexity of unigram, bigram and trigram models trained on one year of New York Times articles **Words (Lexical category)** - Count of each word in the sentence. Numbers and punctuations were removed

Table 5.7: Features for identifying general versus specific sentences

sentiment was often expressed, providing some qualification about a person or event. In the general sentences in Table 5.2, we see for example the phrases "publishing sensation", "very slowly—if at all", "is significant". In a sense, general sentences appear to be more surprising, and evoke in the mind of the reader questions about some missing information or explanation. Specific sentences, on the other hand, are characterized by the use of proper names and numbers.

The list of features is summarized in Table 5.7. Some of our features require syntax information. For sentences from WSJ articles, we compute these features using the manual parse annotations from the Penn Treebank corpus [100]. For other corpora, we obtained the parses using the Stanford parser [75]. We call all features except words as the non-lexical category.

5.4 Classification experiments

In this section, we describe the performance of classifiers built on the data that we collected from both discourse relations and Mechanical Turk annotations. We also validate the proxy data from discourse relations by comparing how the accuracy of a classifier trained on the discourse relations data performs when applied to the direct annotations that we collected from people.

5.4.1 Three types of classifiers

On the discourse relations data, we built two classifiers for distinguishing general and specific sentences: one trained on sentences from Instantiation relations, and one on sentences from Specifications. We built a separate classifier on the examples collected from direct annotation. Each classifier was trained and tested on sentences from the same source. For example, we train a classifier on the general and specific sentences collected from the Instantiations data and test its accuracy on a held-out set of sentences, also from Instantiation relations.

We train a logistic regression classifier[20] with each set of features described above and evaluate the predictions using 10-fold cross validation. For the Instantiations and Specifications data, the general-specific categories have equal number of sentences and the baseline random accuracy is 50%. For the Mechanical Turk annotations, the majority class baseline (specific) is 56%. Table 5.8 shows the accuracy of our features.

The classifier on the turker data, despite having fewer training examples is overall the best performing. With the non-lexical features all put together, the accuracy is 79%. Using only lexical features gives worse results, 71% on this data. The system trained on Instantiations examples is also promising with 75% accuracy for both lexical and non-lexical features. Lexical features are less sparse on larger data and this could be contributing to better performance of these features on the Instantiations data.

The Specifications data however obtains much lower performance, the best accuracy is only 12% above the baseline. It is possible that in Specification relations, the specificity of the second sentence is only relative to that of the first. On the other hand, for

[20] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

Features	Instantiations	Specifications	Turk annotations
NE+CD	68.6	56.1	73.0
language models	65.8	55.7	71.1
word specificity	63.6	57.2	70.2
syntax	63.3	57.3	69.4
polarity	63.0	53.4	67.9
sentence length	54.0	57.2	56.6
all non-lexical	75.0	62.0	79.4
lexical (words)	74.8	59.1	71.5
all features	75.9	59.5	78.2

Table 5.8: Accuracy of different features for classifying general versus specific sentences

Instantiations, there are individual characteristics related to the generality or specificity of sentences. We further verify the suitability of the discourse relations data for this task in the next section.

Among the non-lexical features, the NE+CD class is the strongest with an accuracy of 68% for Instantiations and 73% on manual annotations. Language models, syntax, polarity and specificity features also outperform the baseline by about 15% accuracy. The sentence length features are the least indicative. The non-lexical feature classes though not that strong individually, combine to give the same performance as the word features. The combination of lexical and non-lexical categories does not outperform the accuracies obtained by each individual category.

5.4.2 Validating the use of discourse relations to create examples

We have assumed from the definitions of Instantiation and Specification relations, that their first sentences (*Sent1*) are general and their second (*Sent2*) specific. Further, we used these two sentences independently in two different classes. Now we test this intuition directly. We seek to answer two questions:

- Would people given only one of these sentences in isolation, give it the same judge-

	Instantiations data	
	General	**Specific**
Sent1	29 L5(14), L4(9), L3(6)	3 L5(1), L4(1), L3(1)
Sent2	6 L5(1), L4(3), L3(2)	26 L5(13), L4(9), L3(4)
	Specifications data	
	General	**Specific**
Sent1	10 L5(4), L4(3), L3(3)	6 L5(1), L4(1), L3(4)
Sent2	8 L5(5), L4(3), L3(0)	8 L5(5), L4(2), L3(1)

Table 5.9: Annotator judgements of general/specific nature for Instantiation and Specification sentences

ment of generality as we have assumed?

- How well does a classifier trained on the discourse relations data perform on the direct annotations obtained through Mechanical Turk?

To answer the first question, we included sentences from Instantiation and Specification relations in the dataset for turk annotations. There were 32 Instantiations and 16 Specification relations in the three WSJ articles we annotated and each of these relations is associated with two sentences, *Sent1* and *Sent2*.

In Table 5.9, we provide the annotator judgements and agreement levels on these sentences. The number of sentences x in each category with a certain level of agreement y is indicated as Ly(x). So L5(3) means that three sentences had full agreement 5.

For Instantiations, we find that the majority of *Sent1* are judged as general and the majority of *Sent2* are specific, 80% in each case. But for both *Sent1* and *Sent2*, there is one sentence which all the annotators agreed should be in the opposite class than assumed.

		Accuracy		
Data	**No. examples**	**All features**	**Non-lexical**	**Lexical**
WSJ	293	73.7	76.7	71.6
AQUAINT	287	59.2	81.1	67.5
NYT-science	305	67.2	74.4	58.3

Table 5.10: Accuracies of the Instantiations-trained classifier on the Mechanical Turk annotations

So there are some cases where without context, the judgement can be rather different. But such examples are infrequent in the Instantiation sentences. Hence this dataset closely approximates the general-specific distinction which we wished to learn.

On the other hand, Specifications show a weaker pattern. For *Sent1*, still a majority (62.5%) of the sentences are called as general. However, for *Sent2*, the examples are equally split between general and specific categories. Hence it is not surprising that the Instantiation sentences have more detectable properties associated with the first general sentence and the second specific sentence and the classifier trained with these examples obtains better performance compared with training on Specifications.

Therefore the Instantiations examples appear to be reliable data for our task while Specifications relations do not appear useful for the binary distinction we make in this work. So we further test the validity of the Instantations data by training a classifier on the Instantiations examples and testing it on the annotations obtained directly through Mechanical Turk. High performance on this task would indicate that the Instantiations data while still a proxy provides a similar distinction as that given by people's ratings.

Table 5.10 shows the results for this task. A classifier was trained on the Instantiations data and tested on each of the three sets of annotations from WSJ, AQUAINT and NYT. Since there were sentences in the WSJ test data which overlapped with our Instantiations training set, we removed the overlapping sentences and retrained our classifier while testing on the WSJ data. We find that the Instantiations based classifier has the same accuracy on the directly annotated data compared to when tested on a held-out sample of Instantiations sentences. The highest accuracies are obtained using the non-

Feature set	Accuracy
Nonlexical	72.7
Words	72.1
All features	74.2

Table 5.11: Accuracy on combined set of Instantiations and manually annotated data

lexical features similar to our findings in the previous section. For this feature set, the accuracies are around 75% on the WSJ and NYT data. For the AQUAINT annotations, the accuracy is even higher 81%. While using the word or all features the accuracies are not as high probably due to varying lexical items present on the WSJ corpus compared to other corpora. Accordingly, the word-based classifiers accuracy is 71% on the WSJ data but only 58% on the NYT. The non-lexical features on the other hand, have similar high accuracies on different corpora.

These experiments validate that the Instantiations examples provide a suitable and useful dataset for the general/specific distinction.

5.4.3 Combined classifier

Since both Instantiations and the direct annotations gave good accuracies, we also combined them to obtain a larger set of examples. Here the total general sentences is 1,768 and there are 1,858 specific sentences. So the distribution is almost equal (49% general and 51% specific) and the baseline random performance would be 50% accuracy. The 10-fold cross validation accuracies from non-lexical, word and 'all features' on this full set are shown in Table 5.11. The best accuracy was obtained by combining all features, 74%. Individually, the lexical and non-lexical categories each give 72%.

Since our classification approach has sufficient training data and good accuracy of 75% we used it to analyze writing quality for two genres: summarization and science journalism. Before discussing these we provide further analysis on the manual annotations which helped us obtain a score for specificity rather than binary prediction.

5.5 Graded measure of specificity

Our annotation results suggested that some sentences are harder for people to annotate and others were easy. To understand the relationship better, we analyzed how our classifier handles examples with different agreement levels.

5.5.1 Prediction on examples with different agreement

Table 5.12 gives the accuracy of the Instantiations trained classifier on the turk annotations as in the previous section but also splits the results for examples with different agreement levels. 'Agreement $3+4+5$' indicates all examples with majority agreement, 'Agreement $4+5$' indicates only examples with an agreement level of 4 or 5 and so on.

For all feature sets and test corpora, the accuracies increase steadily as examples with higher agreement are considered. There is at least 10% better accuracy on the examples with agreement 5 compared to all examples that have majority decision. Particularly, for our best-performing feature class "non-lexical", the accuracies are 75% on all examples combined and above 90% on the agreement 5 examples. Therefore the sentences with greater annotator agreement appear to be more clear-cut and easy examples to classify. Note that the agreement levels are not available to the classifier, it was trained on the Instantiations examples.

To study the relationship between classification accuracy and annotator agreement on an example, we further examined the confidence produced from the classifier (logistic regression probability) during prediction. We only used the annotated data from Mechanical Turk for this experiment.

We first combined the predictions for the sentences from the 10 folds in the prediction experiment and split the data into sentences which the classifier predicted correctly (above 0.5 confidence for the right class) and wrong predictions (above 0.5 confidence for the wrong class). Then in each set, we recorded the average value of classifier confidence on examples with different agreement. The results are shown in Table 5.13 for all the data and also when split by corpus. When the mean value in one agreement level is significantly higher (under a two-sided t-test) than at another level, the lower levels are shown within parentheses.

	WSJ sentences				AP sentences			
Examples	**Size**	**All**	**Nonlex**	**Lex**	**Size**	**All**	**Nonlex**	**Lex**
Agreement $3+4+5$	293	73.7	76.7	71.6	287	59.2	81.1	67.5
Agreement $4+5$	198	80.8	88.8	77.7	199	65.8	89.9	74.8
Agreement 5	96	90.6	96.8	84.3	108	69.4	94.4	78.7
	NYT sentences							
Examples	**Size**	**All**	**Nonlex**	**Lex**				
Agreement $3+4+5$	305	67.2	74.4	58.3				
Agreement $4+5$	223	76.4	85.2	66.0				
Agreement 5	82	82.9	92.7	74.4				

Table 5.12: Accuracy of the Instantiations-trained classifier on annotated examples (from Mechanical Turk) split by corpus

We find that when the prediction is correct, the confidence on the examples with highest agreement is on average larger than that on lower agreement examples. For the wrong predictions, we see an opposite trend. On the examples where annotators agreed highly that they belong to one category, the classifier makes lower confidence predictions. On the lower agreement examples, it mispredicts with higher confidence indicating more confusion. We find that for correct predictions, the confidence on examples with agreement level of 4 or 5 is on average higher than that with agreement 3. A two-sided t-test confirmed that this difference was statistically significant. The values in the wrong prediction column are not significantly different in most cases except the AQUAINT data. However, the number of wrong predictions is few overall and the number of mispredicted examples are clearly increasing as the agreement becomes lower. These findings indicate that the classifier performance is related to the annotator agreement or clear general-specific distinction.

Further since the classifier confidence varies according to the annotator agreement these confidence values can be utilized as a measure of graded distinction. We expect that these graded scores will be useful for a variety of tasks as our annotations from

Mechanical Turk clearly show that some examples are easily in one class or another, while some others are harder to place in a binary distinction. The confidence values allow us to capture this aspect.

5.5.2 Score for a text

So far, our predictions were sentence-level. Both the binary class and confidence scores are available only for individual sentences. For text quality prediction, we need to compute the score for an entire article. Simply averaging the specificity scores (confidence values) of sentences might be too coarse since the length of sentences varies greatly. Instead we define a token-level score. We use the classifier to mark for each sentence the confidence for belonging to the *specific* class. We then compute the weighted average of the confidence values of the sentences, where the weights are the number of tokens in each sentence. We call this score *average specificity of words* and use it for many of our analyses.

Below we provide a task-based evaluation of this score.

5.5.3 Task based evaluation: differentiating general and specific summaries

Our aim is to study how well the *average specificity of words* score can capture the specificity of an article as a whole. To do this analysis, we use summaries and source texts from the Document Understanding Conference (DUC) organized by NIST in 2005.[21]

One of the summarization tasks in 2005 was to create summaries that are either general or specific. Input sets for summarization consisted of 25 to 50 news articles on a common topic. *Each input* was also associated with a topic statement which states the user's information need. During the creation of input sets, the annotators were asked to specify for each input, the type of summary that would be appropriate. So annotators provided a desired *summary granularity* for each input: either general or specific. Some example topic specifications and general-specific summary markings are shown in Table 5.14.

There were a total of 50 inputs, 24 of them were marked for general summaries, the remaining for specific.

[21] `http://duc.nist.gov/duc2005/`

Agree	Correct		Wrong	
	No. examples	Confidence	No. examples	Confidence
		All data		
5	277	0.87 (4,3)	9	0.65
4	269	0.81 (3)	45	0.69
3	163	0.74	122	0.71
		WSJ		
5	96	0.87 (4,3)	0	0.0
4	89	0.78	13	0.67
3	52	0.75	43	0.70
		AP		
5	105	0.88 (4,3)	3	0.56 (4,3)
4	81	0.79 (3)	10	0.72
3	54	0.73	34	0.68
		NYT		
5	76	0.85 (3)	6	0.70
4	99	0.84 (3)	22	0.68
3	57	0.74	45	0.73

Table 5.13: The average confidence of the classifier for correct and wrong predictions. The examples are split across the agreement levels and also shown for different subsets of the annotated data. Within parentheses we show the levels whose mean value is significantly less than the value in the column.

General topics	[1] Police deaths: In what manner have police officers died in the line of duty? What are the circumstances surrounding these deaths? [2] Wildlife in Danger of Extinction: What general categories of wildlife are in danger of extinction world-wide and what is the nature of programs for their protection?
Specific topics	[1] Women in Parliaments: Provide information on numbers of women in parliaments across the world, the gap in political power between the sexes, and efforts that have been made to raise the percentages of women in legislative bodies. [2] New Hydroelectric Projects: What hydroelectric projects are planned or in progress and what problems are associated with them?

Table 5.14: Example topic statements for inputs from DUC 2005 summarization task and the type of summary desired for each input

Next gold standard summaries were created by human assessors for all these inputs. A length limit of 250 words is enforced. The assessors are retired information analysts who are experts at summary creation. They were given the input texts and topic statements with the following instructions to create a general or a specific summary as noted.

> A **specific summary** should describe and name specific events (eg. "the bombing of the Pan Am jet over Lockerbie in 1988"), people (eg. "Gadhafi"), places (eg. "Lockerbie"), etc. These specifics are central to the summary and should be generalized only if there is not enough space to include them all.
>
> A **general summary** refers to categories/types of things (eg. "terrorist bombings", "dictators in the Middle East", "Scottish cities") but can refer to specific events, people, places, etc., as illustrative examples if space allows; however, unless the topic statement explicitly requests somthing specific, these examples themselves are not the focus of the summary.

For some inputs (20), 9 summaries each were provided by the assessors, other inputs had 4 summaries. Considering the granularity of inputs, there is a roughly equal distribution of general (146) and specific (154) summaries created by the assessors.

Text	General category	Specific category
Summaries	0.46 (0.13)	0.56 (0.13)
Inputs	0.58 (0.05)	0.60 (0.04)

Table 5.15: Mean value (and standard deviation) of specificity score for inputs and human-written summaries from DUC 2005

We now test if our classifier predictions can distinguish between these general and specific summaries where people relied on an intuitive idea of general and specific content overall in the summary.

We compute the *average specificity of words* score for each summary. The statistics for this score in the general and specific categories are shown in Table 5.15.

For specific summaries, the mean specificity is 0.56, while for general ones it is only 0.46. The difference is also statistically significant under a two sided t-test (p-value of 2.9e-10). This result shows that our predictions are able to distinguish the two types of summaries.

We also computed the specificity scores for inputs in the same manner. Here the mean value is around 0.58 and does not vary significantly between the two classes. So while the inputs do not vary in specificity for the two categories, the summary authors appear to have injected the required granularity during summary creation. This difference in the specificity of summaries is captured by our score.

With these validations of our classifier and its scores, we move on to using these predictions to perform text quality evaluations on a large scale. We first discuss experiments on summarization data and later on our science journalism corpus. No manual annotation of general and specific nature was done during these experiments. We use the classifier trained on the Instantiation sentences for both tasks below. We used the Instantiations data because some of the sentences from the Mechanical Turk data overlap with the science journalism and summarization texts which we wish to analyze. Since the Instantiations based classifier had highly accurate predictions on the different corpora from turk annotations, we choose to use this data for training. Only the non-lexical features were used for our experiments here with the expectation that these features will be more

appropriate for obtaining predictions on data from different corpora. Our experiments so far show that these features have the best accuracy on different corpora and genre.

5.6 Text quality assessment for summarization

As we discussed in the introduction to this chapter, the general-specific nature of texts has been noticed in different studies on summaries. These studies however, involved manual annotation. In this section, we first present experiments where we examined different types of summaries—abstracts and extracts written by people and summaries produced by automatic systems to understand if these texts show differences in specificity. Then we study the relationship between the specificity scores and summary quality in various settings.

5.6.1 Analyzing the abstract creation process

There are two classes of prior studies on summarization where text specificity was discussed. Haghighi and Vanderwende (2009) [61] and Mason and Charniak (2011) [101] have focused on the relationship between text specificity and summary quality. They note that when content selection step is biased to favor general content, the summaries produced by their systems obtained better scores during manual evaluations. On the other hand, the study by Jing and McKeown (2000) [69] manually analyzed sentences present in summaries written by people and note that people sometimes make sentences from the source more general in summaries. This observation is a relevant one for summarization because summaries have to convey information within the specified word limit. Hence many details cannot be included in summaries. Now we examine both these types of data: human summaries and sentences in human summaries to understand how our automatic predictions for specificity relate to these findings from prior work.

General-specific property of human and automatic summaries

In this experiment, we analyzed the specificity trends in different types of summaries. Specifically, we examine abstracts and extracts written by people, automatic summaries

generated by systems, and the original source documents given for summarization.

We obtained news documents and their summaries from the Document Understanding Conference (DUC) evaluations conducted in 2002. We use the data from 2002 because they contain the three different types of summaries we wish to analyze—abstracts and extracts produced by people, and automatic summaries. For extracts, the person could only select complete sentences, without any modification, from the input articles. When writing abstracts people were free to write the summary in their own words.

We use data from the generic multi-document summarization task. There were 59 input sets, each containing 5 to 15 news documents on a topic. The task is to provide a 200 word summary. Two human-written abstracts and two extracts were produced for each input and they were created by trained assessors at NIST. Nine automatic systems participated in the conference that year and we have 524 automatic summaries overall.

For each text—input, human abstract, human extract and automatic summary—we compute a measure of specificity, the *average specificity of words* score which we introduced in Section 5.5.2. The histogram of this measure for each type of text is shown in Figure 5.1.

For inputs, the average specificity of words ranges between 50 to 80% with a mean value of 65%. So news articles tend to have more specific content than generic but the distribution is not highly skewed towards either of the extreme ends.

The remaining three graphs in Figure 5.1 represent the amount of specific content in summaries for the same inputs. Human abstracts, in contrast to the inputs, are spread over a wider range of specificity levels. Some abstracts have as low as 40% specificity and a few actually score over 80%. However, the sharper contrast with inputs comes from the large number of abstracts that have 40 to 60% specificity. This trend indicates that abstracts contain more general content compared to inputs. An unpaired two-sided t-test between the specificity values of inputs and abstracts confirmed that abstracts have significantly lower specificity. The mean value for abstracts is 62% while for inputs it is 65%.

The results of the analysis are opposite for human extracts and system summaries. The mean specificity value for human extracts is 72%, 10% higher compared to abstractive

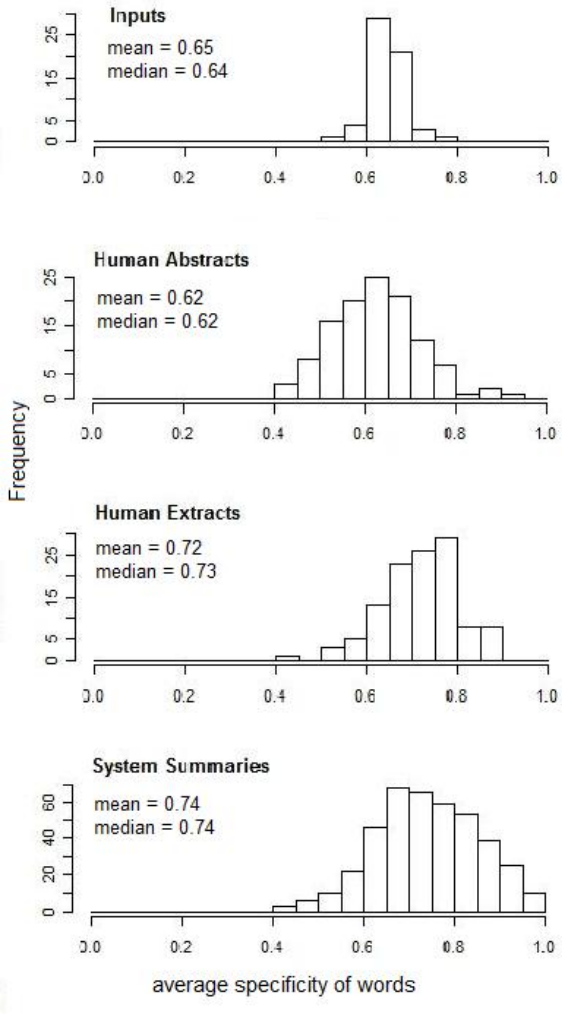

Figure 5.1: Specific content in inputs and human and automatic summaries

summaries for the same inputs. This difference is also statistically significant. System-produced summaries also show a similar trend as extracts but are even more heavily biased towards specific content. There are even examples of automatic summaries where the specificity level reaches 100%. The mean specificity value is 74% which turned out significantly higher than all other types of texts, inputs and both types of human summaries. So system summaries appear to be overwhelmingly specific.

The first surprising result is the opposite characteristics of human abstracts and extracts. While abstracts tend to be more general compared to the input texts, extracts are more specific. Even though both types of summaries were produced by people, we see that the summarization method deeply influences the nature of the summary content. The task of creating extractive summaries biases towards more specific content. So it is obvious that systems which mainly use extractive techniques would also create very specific summaries.

However both types of human summaries have much lower specificity compared to system produced summaries. These findings indicate that for people, summaries involve general information as we had expected. However, we find that current automatic systems have much more specific information. Given our hypothesis about the need for general information in summaries and findings from Haghighi and Vanderwende (2009) and Mason and Charniak (2011), we can expect that the specificity scores could have a direct relationship with the quality of summaries.

General-specific property of compressions

In this experiment, we study the specificity of summaries at sentence-level. Several sentences in summaries can be mapped back to a close source document sentence that conveys similar content. Such mappings have been used as the data for compression tasks, which aim to compress a source sentence to its form in the abstract. We used these mappings on a corpus of human abstracts and source documents to study the specificity of the source and abstract sentence in a mapped pair.

We use the mappings created on the Ziff Davis corpus [63] which contains articles on computer-related products. Several prior studies [52, 76, 103] have used the mappings

Type	Total	% total	Avg deletions	Avg subs.	Orig length	Compr. rate
SS	6371	39.9	16.3	3.9	33.4	56.6
SG	5679	35.6	21.4	3.7	33.5	40.8
GG	3562	22.3	9.3	3.3	21.5	60.8
GS	352	2.2	8.4	4.0	22.7	66.0

Table 5.16: Specificity predictions on paired source and abstract sentences

on this data for compression experiments. The mapped (alignment) pairs are produced by allowing a limited number of edit operations to match a source sentence to one in the abstract.

We use the alignments created by Galley and McKeown (2007) [52] who allowed *any* number of deletions and upto 7 substitutions. There are 15964 such pairs in this data.

We ran the classifier individually on each source sentence and abstract sentence in this corpus. Then we counted the number of pairs which undergo each transformation such as general-general, general-specific from the source to an abstract sentence. These results are reported in Table 5.16.

We find that during compression, frequent transformations are specific-specific (SS) and specific-general (SG). Together they constitute 75% of all transformations. But for our analysis, the SG transformation is most interesting. One third of the sentences in this data are converted from originally specific content to being general in the abstracts.

The table also provides the average number of deletion and substitution operations associated with sentence pairs in that category as well as the length of the uncompressed sentence and the compression rate. Compression rate is defined as the ratio between the length in words of the compressed sentence and the length of the uncompressed sentence. So lower compression rates indicate greater compression. We find that the SG transition has the highest compression. The original source sentences for the SG transitions are long but there is considerable deletion to create a general from specific sentence (highest value of 21 compared to average deletions of 16 and lower for the other transition types). Table 5.17 shows some sentence pairs which involve specific to general transformation.

This result again indicates that the general-specific distinction is a highly useful one

[1] American Mitac offers free technical support for one year at a toll-free number from 7:30 to 5:30 P.S.T.
American Mitac offers toll-free technical support for one year.
[2] In addition to Yurman, several other government officials have served on the steering committee that formed the group.
Several government officials also served on the steering committee.
[3] All version of the new tape drives, which, according to Goldbach, offer the lowest cost per megabyte for HSC-based 8mm tape storage, are available within 30 days of order.
The products are available within 30 days of order.

Table 5.17: Example specific to general (in italics) compressions

for summarization. In the compression task, for example, standardly only importance of words retained and the grammatical correctness of the sentence are considered as objectives to optimize. Since several of these sentences are made more general during compression, we believe that our classifier and scores can help create better compressions by providing a way to incorporate a desired level of specificity for the compressed sentence.

5.6.2 Relationship to summary quality

Here we directly study the relationship between the specificity of summaries and their content and linguistic quality scores. We do this analysis on system produced summaries since they have greater variability in scores allowing us to examine how specificity varies on summaries with different perceived quality. Moreover these findings on system generated text could be directly useful for system development.

We present three studies in this section: relationship to content quality, relationship to linguistic quality and relationship to quality of general-specific summaries.

Content quality

We used data from the generic multi-document summarization task at DUC 2002 which we used for the summary analysis in the previous section. There are a total of 524 system summaries.

Each summary was evaluated by human judges for content and linguistic quality during the DUC evaluation. The quality of content was assessed in 2002 by means of a COVERAGE score. The coverage score reflects the similarity between content chosen in a system summary and that which is present in a human-written summary for the same input. The human-written abstracts are produced by trained assessors at NIST. One human abstract is chosen as the reference. It is divided into clauses and for each of these clauses, judges decide how well it is expressed by the system produced summary (as a percentage value). The average extent to which the system summary expresses the clauses of the human summary is considered as the coverage score. These scores range between 0 and 1.

We computed the Pearson correlation between the specificity of a summary (as described in Section 5.5.2) and its coverage score, and obtained a value of -0.16. The correlation is not very high but it is significant (pvalue 0.0006). Therefore specificity appears to be related to content quality and more specific content is indicative of lower scores.

However, specificity is only indicative of how the content is expressed and is more or less independent of the importance of the content itself. Two summaries can have the same level of specificity but vary in terms of the importance of the content present. In order to control for the importance of content, we tested adding specificity and importance scores as predictors of content quality.

For content importance, we compute ROUGE, the standard approach for automatic evaluation of summary content. ROUGE [86, 87] is a suite of tools to compute n-gram overlap measures between human abstracts and system summaries. These overlap scores have been shown to correlate highly with human judgements of similarity between the system summary and reference. We use the same reference as used for the official coverage score evaluation and compute ROUGE-2 which is the recall of bigrams of the human summary by the system summary. Next we train a regression model on our data using the ROUGE-2 score and specificity as predictors of the content coverage score. We then inspected the weights learnt in the regression model to identify the influence of the predictors. Table 5.18 shows the mean values and standard deviation of the beta coefficients. We also report the results from a test to determine if the beta coefficient for a particular

Predictor	Mean β	Stdev. β	t value	p-value
(Intercept)	0.212	0.03	6.87	2.3e-11 *
rouge2	1.299	0.11	11.74	$<$ 2e-16 *
avgspec	-0.166	0.04	-4.21	3.1e-05 *

Table 5.18: Results from regression test for predicting content coverage scores using ROUGE and specificity values

predictor could be set to zero. The p-value for rejection of this hypothesis is shown in the last column and the test statistic is shown as the 't value'. We used the *lm* function in the R toolkit[22] to perform the regression.

From the table, we see that both ROUGE-2 and average specificity of words (avgspec) turn out as significant predictors of summary quality. But the R-squared value is only 0.275. Other factors such as the difficulty of the input text (from the point of view of creating a summary) are also known to influence summary quality scores [113].

Relevant content is highly important as shown by the positive beta coefficient for ROUGE-2. At the same time, good summaries are associated with low specificity, a negative value is assigned to the coefficient for this predictor.

Linguistic quality

We have seen from the above results that lower specificity is associated with higher content quality. A related question is the relationship between specificity and the linguistic quality of a summary. We briefly examine this aspect here.

In DUC 2002, linguistic quality scores were only recorded as the number of errors in a summary, not a holistic score. Moreover, it was specified as a range—errors between 1 and 5 receive the same score. So we use another dataset for this analysis. We use the system summaries and their linguistic quality scores from the TAC 2009 query focused summarization task[23]. In this dataset, each summary was manually judged by NIST assessors and assigned a score between 1 to 10 to reflect how clear it is to read. The score

[22] http://www.r-project.org/

[23] http://www.nist.gov/tac/2009/Summarization/update.summ.09.guidelines.html

linguistic quality score	no. of summaries	avg specificity
1, 2	202	0.71
5	400	0.72
9, 10	79	0.77

Table 5.19: Number of summaries at extreme levels of linguistic quality scores and their average specificity values

combines multiple aspects of linguistic quality such as clarity of references, amount of redundancy, grammaticality and coherence.

Since these scores are on an integer scale, we do not compute correlations. Rather we study the specificity, computed in the same manner as described previously, of summaries at different score levels. Here there were 44 inputs and 55 systems. In Table 5.19, we show the number of summaries and their average specificity for 3 representative score levels—best quality (9 or 10), worst (1 or 2) and mediocre (5). We only used summaries with more than 2 sentences as it may not be reasonable to compare the linguistic quality of summaries of very short lengths.

Summaries with greater score have a higher level of specificity. The summaries with average to low scores (1,2,5 on the scale) do not have noticeable differences in specificity. However, the specificity of the best summaries (9, 10) are significantly higher than that with medium and low scores (two-sided t-test). The finding indicates that opposite to our results with content quality, good summaries are associated with higher specificity levels when we consider the linguistic quality dimension.

As an example, consider the summary in Table 5.20. This summary has a low specificity value of 0.45 and its linguistic quality score is 1. This summary as a whole appears uncontentful and difficult to read. One reason for the low quality could be that the general sentences in the summary do not have any substantiation, such as the first and last sentences. General sentences cannot stand alone and need adequate support and details. But currently, very few systems even make an attempt to organize their summaries. When such overly general content and general content without proper context is present, it appears that the summaries are associated with low linguistic quality scores.

"We are quite a ways from that, actually." As ice and snow at the poles melt, the loss of their reflective surfaces leads to exposed land and water absorbing more heat. It is in the middle of an area whose population –and electricity demands–are growing. It was from that municipal utility framework, city and school officials say, that the dormitory project took root. "We could offer such a plan in Houston next year if we find customer demand, but" "we haven't gone to the expense of marketing the plan. We get no answers."

Table 5.20: Example general summary with poor linguistic quality

We see that specificity is related to both content and linguistic quality of summaries though in opposite directions.

Which general sentences are useful?

We find that overall texts whether input or summaries, have more specific content than general. However more general content is present in summaries which received greater scores from human judges. Therefore we can expect that certain types of general sentences could be useful for inclusion in summaries.

Here we provide a preliminary analysis of general sentences that were chosen to be included in extractive summaries *created by people*. These sentences can provide an understanding of the types of general sentences considered as useful by people for their summaries. We show in Table 5.21, the ten extract sentences that were predicted to be general with highest confidence. The first sentence has a 0.96 confidence level, the last sentence has 0.81.

These statements definitely create expectation and need further details to be included. Taken out of context, these sentences do not appear very contentful. However despite the length restriction while creating summaries, humans tend to include these general sentences. Table 5.22 shows the full extract which contains one of the general sentences ([9] "Instead it sank like the Bismarck.").

When considered in the context of the extract, we see clearly the role of this general sentence. It introduces the topic of opposition to Bush's nomination for a defense secretary. Moreover, it provides a comparison between the ease with which such a proposition could have been accepted and the strikingly opposite situation that arose—the

[1] Folksy was an understatement.
[2] ”Long live democracy”!
[3] The dogs are frequent winners in best of breed and best of show categories.
[4] Go to court.
[5] Tajikistan was hit most hard.
[6] Some critics have said the 16-inch guns are outmoded and dangerous
[7] Details of Maxwell’s death are sketchy.
[8] ”Several thousands of people who were in the shelters and the tens of thousands of people who evacuated inland were potential victims of injury and death”.
[9] Instead it sank like the Bismarck.
[10] ”The buildings that collapsed did so because of a combination of two things: very poor soil and very poor structural design,” said Peter I. Yanev, chairman of EQE Inc., a structural engineering firm in San Francisco.

Table 5.21: Example general sentences in humans extracts

overwhelming rejection of the candidate by the senate. So sentence [9] plays the role of a topic sentence. It conveys the main point the author wishes to make in the summary and further details follow this sentence.

But given current content selection methods, such sentences would rank very low for inclusion into summaries. So the prediction of general sentences could prove a valuable task enabling systems to select good topic sentences for their summaries. However, proper ordering of sentences will be necessary to convey the right impact but this approach could be a first step towards creating summaries that have an overall theme rather than just the selection of sentences with important content.

We also noticed some other patterns in the general sentences chosen for extracts. A crude categorization was performed on the 75 sentences predicted with confidence above 0.65 and are shown below:

first sentence : 6 (0.08)

last sentence : 13 (0.17)

comparisons : 4 (0.05)

Summary d118i-f:

- President-elect Bush designated Tower as his defense secretary on Dec. 16. [Specific]
- Tower's qualifications for the job –intelligence, patriotism and past chairmanship of the Armed Services Committee –the nomination should have sailed through with flying colors. [Specific]
- *Instead it sank like the Bismarck. [General]*
- In written testimony to the Senate panel on Jan. 26, Tower said he could "recall no" actions in connection with any defense activities" in connection with his work for the US subsidiary [Specific]
- Tower has acknowledged that he drank excessively in the 1970s, but says he has reduced his intake to wine with dinner. [General]
- The Democratic-controlled Senate today rejected the nomination of former Texas Sen. John Tower as defense secretary, delivering a major rebuke to President Bush just 49 days into his term. [Specific]
- The Senate's 53-47 vote came after a bitter and divisive debate focused on Tower's drinking habits, behavior toward women and his business dealings with defense contractors. [General]

Table 5.22: Example extract with a general sentence from Table 5.21

attributions : 14 (0.18)

A significant fraction of these general sentences (25%) were used in the extracts to start and end the summary, likely positions for topic sentences. Some of these (5%) involve comparisons. We detected these sentences by looking for the presence of connectives such as "but", "however" and "although". The most overwhelming pattern is presence of quotations, covering 18% of the sentences we examined. These quotations were identified using the words "say", "says", "said" and the presence of quotes. We can also see that three of the top 10 general sentences in Table 5.21 are quotes.

We expect that in future work, we will develop techniques to identify and place useful general sentences in summaries and examine how people perceive these compared to the standard optimization for specific information carried out by most systems.

Quality of specialized summaries

So far, we examined the effect of specificity on the quality of generic summaries. Now, we examine whether this aspect is related to the quality of summaries when they are optimized to be either general or specific content. We perform this analysis on DUC 2005[24] data where the task was to create a general summary for certain inputs. The data set was introduced in the task based evaluation detailed in Section 5.5.3. However, in that section we used the human summaries. Here we use the summaries produced by systems for the same task and their evaluations computed by NIST assessors.

We tested whether the degree of specificity (computed as the *average specificity of words*) is related to the content scores[25] of system summaries of these two types—general and specific. The Pearson correlation values are shown in Table 5.23. Here we find that for specific summaries, the level of specificity is significantly positively correlated with content scores. For the general summaries there is no relationship between specificity and content quality.

These results show that specificity scores are not consistently predictive of distinctions within the *same* class of summaries. Within general summaries, the level of generality is not related to the scores obtained by them. However, for specific summaries the specificity

[24] `http://duc.nist.gov/duc2005/`

[25] We use the official scores computed using the Pyramid evaluation method [114]

Summaries	correlation	p-value
DUC 2005 general	-0.03	0.53
DUC 2005 specific	0.18*	0.004

Table 5.23: Correlations between content scores and specificity for general and specific type automatic summaries in DUC 2005

score is significantly positively correlated with the summary score. We also computed the regression models for these two sets of summaries with ROUGE scores and specificity, and specificity level was not a significant predictor of content scores. These findings could indicate that within a homogeneous class, when either all summaries are general or all are specific, then the level of general and specific nature has less impact on the quality of summaries. In other words, a certain level of general or specific nature could characterize these summaries but above that further general or specific nature is not indicative of summary quality.

5.7 Text quality assessment for science journalism

Now we turn to experiments on text quality prediction for science journalism based on the general-specific distinction. We obtained the predictions from the classifier for each sentence in our science journalism corpus and composed several features to indicate specificity scores at article level.

Overall specificity features: The specificity score explained in Section 5.5.2 is added as a feature (AVGSPECW). We also include the fraction of specific sentences in the article (SPEC_SENT). We also obtain the confidence of each sentence belonging to the 'specific' class and compute the mean (SPEC_MEAN) and variance (SPEC_VAR) of this confidence measure as features.

Sequence features: We added as features the proportion of different transitions between adjacent sentences (GG, GS, SG, SS where G indicates general sentence and S indicates specific) out of the total transitions. We also measured the sizes of contiguous blocks of general and specific sentences. We group the blocks into three bins depending on

feature	mean VERY GOOD	mean TYPICAL
Higher value in VERY GOOD		
GGProp	0.36	0.34
GL	0.18	0.17
Higher value in TYPICAL		
perspec	0.41	0.43
varspec	0.05	0.052
avgspecw	0.54	0.56
SSProp	0.19	0.20
SL	0.09	0.10

Table 5.24: Mean values of specificity features for the quality categories on science news. Only those features where the mean value was significantly (95% confidence level) different between the categories is reported.

block size: sizes of 1, 2 and above 3. The proportion of blocks that fell in each category were added as features. These features are indicated as S1, S2, SL, G1, G2, GL, where L indicates block size above 2.

We first tested how these features vary between GOOD and TYPICAL writing using a random sample of 1000 articles taken from the VERY GOOD category and another 1000 taken from the TYPICAL category. No pairing information (based on topic) was used during this sampling as we wanted to test overall if these features are indicative of good articles rather than their variation within a particular topic. A two sided t-test was computed to test if the mean value of a feature varied significantly between the two classes. The results are shown in Table 5.24.

Seven features are significantly different between the categories. The GGProp and GL blocks have higher values in the GOOD writing. SSProp, SL and a number of specificity scores are higher in the TYPICAL class. Both trends indicate that better written articles are associated with more general content than the average articles similar to our findings on the summary analysis tasks.

The features were then input to a classifier in the two setups that we introduced in

Features	Any topic	Same topic
Specificity features	56.2	54.6

Table 5.25: Accuracy of specificity features for predicting quality of science news articles

Section 3.5. The features for the pairwise setting (same-topic) are the difference in feature values for the two articles. A random baseline would be accurate 50% of the time. We performed 10-fold cross validation over our dataset using a SVM classifier. We used a radial basis kernel and tuned the regularization and kernel parameters using cross validation on the development data.

The accuracy using our features is 56.2% for the any-topic setup and slightly lower 54.6% for comparing articles with the same topic. These accuracies are significantly above chance considering the large number of test examples. However they are still low for a text quality prediction task. However, since these features show significant differences between the article categories in our corpus, we expect that they will augment other features relevant for text quality and provide improved accuracies when combined with them.

5.8 Related work

In this section we review theories that are related to the general-specific distinction. As we defined in Section 5.1, we consider specificity to represent the degree of detail. Some content is detailed, others only provide a topic statement. The topic statements also have the property that a reader would need further information from other parts of the article in order to fully comprehend the statement. In our work we performed a binary distinction between these two types of sentences.

A closely related idea is work on *granularity* [109] which we describe in Section 5.8.1. Granularity assumes that coarse or high level facts in the text are composed of other lower level facts and therefore can be broken down. The difference between coarse and fine-grained facts is based on amount of detail just as in our general-specific classification. However, granularity as defined in prior work appears to focus on the relative difference

in detail between two units. In contrast, our work is designed to predict for individual sentences or text units whether it is a high level statement or detail.

In Section 5.8.2, we contrast our work with the notion of genericity. *Genericity* differentiates individual from a group or on the other hand, a specific irregular event from one that is habitual. For example, "John plays the violin." is habitual versus "John gave an amazing performance with his violin yesterday." is a specific event. These distinctions are less directly related to our idea of specificity based on the details in the text.

Finally, in Section 5.8.3, we discuss work that is related to our use of discourse relations for studying the general-specific distinction. We used the Instantiation discourse relations from the Penn Discourse Treebank to obtain examples for general and specific sentences. We discuss in this section, how the idea of general and specific sentences also exist in discourse relations from discourse frameworks such as the Segmented Discourse Representation Theory (SDRT) [3] and the Rhetorical Structure Theory (RST) [97].

5.8.1 Granularity

The idea of granularity focuses on the relationship between topical facts and low level facts in a text. A study of granularity shifts in texts was done recently in Mulkar and Hobbs (2011) [109]. Specifically, they assume that high or coarse level facts are made up of finer level ones which compose to make up the coarse fact. Further it is assumed that both levels of facts, a high level and some of its associated finer facts are often present explicitly in the text. The idea is similar to our definition that in well-written texts, general sentences are substantiated with relevant specific details.

Mulkar and Hobbs [109] propose that the substantiation or finer level detail is related to the coarse fact typically though three relations:

- The entities of fine-grained fact have a part-of relation with those of the coarse fact.
- The event of the fine-grained fact has a part-of relation with that of the coarse fact.
- The fine-grained event is a 'cause' of the higher level event.

An example from their paper is given below.

> The San Francisco 49ers moved ahead 7-3 11 minutes into the game when William Floyd scored a two-yard touchdown run.

This sentence can be divided into the two clauses below and they have different granularities.

Coarse level: The San Francisco 49ers moved ahead 7-3 11 minutes into the game.

Fine level: William Floyd scored a two-yard touchdown run.

The fact that the San Francisco 49ers team moved ahead in the game is a high level event. It occurred because a player from the team scored a touchdown run.

We can notice that the different components hypothesized by Mulkar and Hobbs are present in this example.

- William Floyd is part of the San Franciso 49ers.
- A touchdown event is part of the event of moving ahead in the game.
- William Floyd scoring a touchdown causes his team to move ahead in the game.

As such we can notice that the definition is a relative one, and seeks to relate two text units where one is at a coarse level and the other has finer granularity. This difference contrasts with our work where the goal is to individually characterize sentences as general or specific. Individual characterization is particularly useful for assessing writing quality. Analyzing sentences individually can indicate general sentences with missing details and specific details which are missing topic statements. Similarly individually tagging sentences as general or specific allows us to then examine their sequence for writing quality.

In their study, Mulkar and Hobbs collected paragraph pairs where these relations hold and ask people to annotate the granularity of the pair indirectly by answering four questions: 1) whether one paragraph causes another, 2) is more detailed than another, 3) one is a subevent of the other and 4) whether the event in one paragraph happens after another. They found that people agreed highly on their answers to these questions (except causality which was hard for people to interpret). Based on these results, they propose that granularity distinctions are easily noticeable by people and is a property that they will agree on.

But they did not focus on developing automatic ways to identify granularity in text or apply them in applications.

5.8.2 Genericity

Another related notion to general-specific is that of genericity. Linguistic theories [78] describe genericity at least two levels–sentences and entities.

At the sentence level, the differentiation is between an event that is habitual versus that which is episodic.

For example consider the following sentences.

a) The bomb exploded.

b) Bombs explode when ignited.

Sentence (a) is an episodic sentence giving information about a specific bomb while sentence (b) talks about the general property of bombs.

At noun phrase level, the distinction is based on whether the noun phrase describes a class of individuals (generic) versus those which refer to a specific individual. Consider these sentences.

a) The lion is a mammal.

b) The lion at the circus yesterday performed great tricks.

The same noun phrase 'the lion' has different interpretations in the two sentences. In the first, it refers to the class of lions rather than to any specific lion. The same phrase in the second sentence refers to a particular lion in the circus that is being discusssed.

There are studies [102, 139] that have explored how to automatically predict such distinctions and they have been motivated by different end applications. These notions are different from our work in that they are related to the content being conveyed by the author, for example expressing a rule-like event or referring to a class of individuals. They do not refer to the realization of content (in terms of detail) as we have assumed in our work. As a consequence, these ideas are less related to writing quality per se in contrast to the general-specific distinction.

However, there are some interesting connections between the general specific nature and these ideas.

Some of the sentences which we would consider as general can also be seen as a habitual facts. For example

a) Mr. Murakami is a publishing sensation in Japan.

b) The Booker prize always creates controversey.

Similarity, noun phrases which frequently have a generic interpretation, for example, 'animals', 'humans' etc would under our general-specific framework be examples of words having low specificity. As a result, work which aims to predict genericity could also be useful for the distinctions which we seek in our work.

Reiter and Frank [139] present an automatic approach to identify generic noun phrases. They evaluate their classifier on a corpus of generic and specific noun phrases available with the ACE-2 corpus. Mathew and Katz [102] focus on automatically predicting habitual verus episodic sentences. They annotated a sample of Wall Street Journal sentences for this distinction particularly obtaining sentences that have verbs which have high ambiguity between habitual and episodic sense. Both studies built supervised classifiers which utilize a number of verb, tense and syntactic features. Understandably, some findings reported in these work have similarities with the results in our work. For example, both studies report that words in plural number are associated with habitual events or generic noun phrases. In our work, we also found that plural nouns is a significantly useful feature with higher value in the general class of sentences.

5.8.3 Discourse relations and general-specific nature

In our work, we have used discourse relations to obtain proxy data for general and specific sentences. In the introduction to this chapter, we motivated this choice using the definitions of Instantiation and Specification relations from the Penn Discourse Treebank. These two relations are types of the 'Expansion' class of relations annotated in the PDTB. Other discourse frameworks also record the difference in specificity of content of the two sentences or units in certain discourse relations. These relations also fall under the broad class of Elaboration and Expansion relations in the respective frameworks.

In the Segmented Discourse Representation Theory [3], the elaboration relation is defined as introducing a new level in the text, one that introduces extra detail. These levels are hypothesized as providing a hierarchical structure to the text. Similarly, in Rhetorical Structure Theory [97], one subtype of elaboration relation is called the 'generalization: specific' relation. The annotation manual for RST relations created by Marcu [16] defines this relation as follows: The nucleus presents a concept and the satellite defines the concept in more detail. Moreover, it appears that granularity differences may not only be confined to elaboration class of relations. As discussed in Section 5.8.1, Mulkar and Hobbs in their theory of granularity focus on causal relations. Central to their model is a causal relationship between the event at finer granularity and coarse granularity.

These analogous definitions show that the difference between general and specific sentences is easily noticed in discourse and provide greater support for our use of sentences from discourse relations as training data for specificity.

5.9 Future work

Beyond this thesis, we believe that our work opens up interesting questions to tackle in the future. Some of these ideas are related to use of the general-specific distinction in applications and others are avenues for exploring language properties.

5.9.1 Relevance to discourse parsing

We have utilized a hypothesis about discourse relations in order to create data for general and specific sentences. Given our results, we see that our features can successfully separate out these two classes obtained from the discourse relations. At this point, one can also address the opposite task of using the general-specific notion for discourse parsing.

Particularly, in prior work on discourse parsing, researchers have not focused on expansion relations since they are a large class and often considered as "catch-all". In the Penn Discourse Treebank one-third of all explicit relations (a discourse connective such as 'because' or 'but' is present and signals the relation) and more than half of all implicit relations (no discourse connective) are expansion relations. As a result, most discourse parsing work [89, 99, 122, 151] has not focused on identifying features that might be in-

dicative of expansion relations partly because the differences for more semantially salient relations such as contrast and cause relations are better understood compared to expansions.

Similarly, in the RST framework, when relations are computed between large segments such as paragraphs, an elaboration relation is assigned solely based on whether two paragraphs have a high degree of similarity with respect to their words [98].

But our work shows that some of the expansion relations could contain interesting distinctions we can take advantage of. In another study not reported in this thesis [92, 93], we had studied coreference patterns between the arguments of different discourse relations. Here we found that discourse relations vary in how much coreference is present between their arguments. Expansions relations also had the unique property that it had the least degree of coreference. We believe that such distinctions can be added to discourse parsing studies to improve the classification of discourse relations.

In fact, a preliminary study, Howald and Abramson (2012) [64] found that if we group SDRT relations into certain classes depending on granularity, for example, 'elaboration' relations involve increase in granularity between the first sentence and the next, 'result' relation has a decrease in granularity and for 'alternation' and 'narration' relations, the granularity of the two sentences remains the same. They find that using this information helps to predict the discourse relations with better accuracy compared to when granularity was not included as a feature. However, they use predefined granularity tags for the different relations and do not have a way to quantify general-specific nature explicitly. We can hope that our system can bridge this gap and be useful for discourse relation prediction as well.

5.9.2 Relevance to summarization and information retrieval

Our work highlights the fact that adding structure to summaries can create better quality summaries. However, our work only has a preliminary study of which general sentences are preferred and extracted by people for their summaries. In future, we will explore selection and ordering of general and specific information within a summarization system.

Similarly, the idea of general or specific nature can be helpful during information

retrieval. Some audiences may want only a general idea of the search query, others may be interested in specific details. A system to take such preferences into account may be quite useful. Raymond, Lai and Li (2009) [81] present work that aims to rerank documents based on information granularity computed using a domain ontology. They use two ideas to compute specificity. The first metric uses the depth of a term from the domain ontology to approximate terminology specificity. The second idea is that if the document discusses closely related concepts, it would be more specific compared to one that discusses a variety of concepts.

5.10 Conclusions

In this chapter, we presented a new metric for text quality based on the specificity of the text. We showed that people can make the distinction between general and specific sentences fairly well. For two-thirds of our data, either four or all five of our annotators agreed on the class to assign. Moreover, we found that naturally occurring data annotated for discourse relations were also useful to create training data for this task. Our automatic classifier achieves an accuracy of 75% which is suitable and reliable for use in other applications. Based on the success of the classification approach, we showed how the specificity of a text is related to text quality for two genres: automatic summaries and science journalism. For automatic summaries, lower specificity was indicative of content quality. At the same time, the linguistic quality of general summaries was low. For science journalism, both specificity of content and the sequence of general and specific sentences is indicative of the text quality categories on our corpus. For this genre as well, more general content is associated with higher text quality.

Chapter 6

Indicators of reader interest

The writer's passion for the topic drives the writing, making the text lively, expressive and engaging.

Phrasing is original–even memorable–yet the language is never overdone.

Striking variety in structure and length gives writing texture and interest.

Voice, Word choice and Sentence fluency traits (Section 2.1)

Among the traits which are considered essential for good writing, those related to reader interest are the least explored for computational work. In the Six Traits rubric [150], the categories 'voice', 'word choice' and 'sentence fluency' are the ones related to this aspect of quality. They identify texts as interesting and engaging or otherwise dull. Note that a text need not have these aspects and still be error-free, have good organization and content. These reader interest properties determine whether the text is elegant, beautifully written and interesting to read.

In this chapter, we introduce measures for predicting reader interest of articles from the science journalism genre. These experiments use the text quality categories from the science journalism corpus which we introduced in Chapter 3. Several of our features apart from being designed to indicate engaging nature of writing are also specific to the

science news genre. Use of genre-specific measures is little studied in text quality work in the past but there is ample evidence that such features will be helpful.

There are unique patterns of writing which are noticeable for news in general and also science journalism. Therefore features related to the specific writing patterns of a genre could provide a boost over those which are general across genres. Journalism studies refer to patterns in news writing as *news frames*. A news frame is the selection of a particular way of reporting an issue and varies in terms of the type of main content that is presented and how the article is organized and written. The differences are perhaps best understood with examples. For example, in general news a few common frames are (definitions are taken from Sametko and Valkenburgh (2000) [145]):

> Conflict frame: This frame emphasizes conflict between individuals, groups, or institutions as a means of capturing audience interest.
>
> Human interest frame: This frame brings a human face or an emotional angle to the presentation of an event, issue or problem.
>
> Economic consequences frame: This frame reports an event, problem or issue in terms of the consequences it will have economically on an individual, group, institution, region or country.
>
> Morality frame: This frame puts the event, problem, or issue in the context of religious tenets or moral prescriptions.
>
> Responsibility frame: This frame presents an issue or problem in such a way as to attribute responsibility for its cause or solution to either the government or to an individual or group.

Sametko and Valkenburgh performed a large scale analysis of a few thousand newspaper articles and found that the responsibility and conflict frames are widely popular for coverage of political news. In a similar vein, other studies have reported on specialized frames that are used for science journalism. One such study by Nisbet, Brossard and Kroepsch (2003) [116] examined what types of frames were employed for news reporting during the different stages of policy development related to a scientific issue. They focus on the topic of stem cell research. Media attention surrounding such issues varies during

different times and different types of reporting styles are followed by journalists. Nisbet, Brossard and Kroepsch analyze a large collection of New York Times and Washington Post articles on stem cell research to identify trends in news framing. Abridged short descriptions of some of the frames they used are given below.

> New Research: Focus on new stem cell-related research released, discovery announced, new medical or scientific application announced.
>
> Scientific background: Focus on general scientific or medical background of stem cell-related research or applications. Includes description of previous research, recap of "known" results and findings.
>
> Scientific/technical controversy or uncertainty: Focus on scientific uncertainty over efficacy or outcomes of stem cell-related research and applications.
>
> Public opinion: Focus on the latest poll results, reporting of public opinion statistics, general references.
>
> Anecdotal personalization: Focus on a patient, or the families/friends of a patient, who is receiving stem cell-related treatment.

Given that it is so well documented in prior literature that journalists choose and place much emphasis on the style for writing an article, we expect that the science journalism genre is an apt one for studying which properties of the writing are successful in creating engaging articles. It is also found that the presentation of content as different news frames influences readers' thoughts and recall of information about the topic [166].

In this chapter, we develop a system to predict the quality of science news articles. Its diverse feature set involves measures which indicate *interesting* content and writing together with those that have been previously developed to indicate well-written text.

We design and implement measures for six facets of writing that are related to reader interest: 1) use of visual language, 2) involving people in the story, 3) creative and surprising use of language, 4) sub-genre of the article, 5) use of sentiment and emotions, and 6) the amount of explicit research descriptions. We study how these aspects are distributed in the quality categories in our corpus and also their strengths in making a prediction of

the category. Rather than add a large number of features which may be indicative of these dimensions indirectly, we aimed to develop measures which specifically indicate a particular aspect. Otherwise when a feature turns out as a significant indicator of the quality categories, we still may not be able to associate the feature with any particular writing aspect. We also validate a few of our features using human annotations to understand if a text that is ranked high according to a particular feature is also considered by people to have the corresponding property which the feature represents. These annotations gives additional strength to our claims of which aspects are related to text quality in science journalism. Sections 6.1 and 6.2 focus on the development of features and the annotation study performed to understand the representative nature of the features.

We then examine how these features help to predict the quality categories in our corpus. While we have used the intentional structure model and the text specificity features to predict the quality of these articles in the previous chapters (see Sections 4.5 and 5.7), the accuracies that we obtained were quite low. We show that features related to interest which we develop in this chapter are much more predictive of quality differences leading to accuracies of 77% when articles are compared without regard to topic and 70% when comparing articles with the same topic. A detailed analysis of classification accuracy and strengths of different feature classes is presented in Section 6.4.

We also show that the interest and genre-specific features complement those which aim to identify readable and well-written texts. We combine a comprehensive set of features from prior work on readability and well-written nature of articles with those we developed for reader interest and find that all these measures together are necessary for text quality prediction on our corpus. These analyses are presented in Section 6.5.

Finally, we examine the influence of topic and content of the article on its quality. The metadata available in the New York Times corpus allows us to study which topics are most frequently chosen in the GREAT article set. We present experiments on automatic prediction of quality based on features derived from the metadata and also approximate topic information using words in the articles. In Section 6.5.2, we report the results from this experiment and they provide evidence that topic features are also useful indicators of text quality in this genre.

6.1 Facets of writing in science news

We discuss six facets of writing which are easily noticeable in science news articles and which we hypothesized will have an impact on text quality. They are 1) visual nature, 2) people-oriented content, 3) beautiful language, 4) sub-genre of text, 5) sentiment and emotional language, and 6) degree of research content. Several other aspects could also be relevant to quality such as the use of humour, metaphor, suspense and clarity of explanations. We choose the six facets above based on evidence from prior literature for their relationship to quality and also the feasibility of measuring them automatically.

We describe each facet below and also the motivation for proposing it as an indicator for text quality. We also explain how we computed features related to each property and report how these features vary in the VERY GOOD and TYPICAL categories in our corpus. To do this analysis, we randomly sampled 1000 articles from each of the two categories as representative examples. We compute the value of each feature on these articles and use a two-sided t-test to check if the mean value of the feature is higher in one class of articles versus another. A p-value less than 0.05 is taken to indicate significantly different trend for the feature in the VERY GOOD versus TYPICAL articles.

Finally, we also present an annotation study to test whether the features capture the intended facets. We asked annotators to judge whether the texts which rank high or low according to a particular feature value contain the facet that the feature represents. We performed these annotations for a few features taken to represent all of the six facets.

6.1.1 Visual nature of articles

Some texts create an image in the reader's mind. For example, the snippet below has a high visual effect. All the snippets in this chapter are taken from the science journalism corpus which we discussed in Chapter 3.

> When the sea lions approached close, seemingly as curious about us as we were about them, their big brown eyes were encircled by light fur that looked like makeup. One sea lion played with a conch shell as if it were a ball.

Such vivid descriptions engage and entertain readers. The relationship between visual quality and cognition has been well-documented. Kosslyn (1980) [77] found that in several situations people spontaneously form images of concrete words that they hear and use them to answer questions or perform other tasks. Imagery has also been found to help with recall of information. There is evidence that when people visualize the items corresponding to a given pair of words, they recall the pair better compared to when they do not create mental images [12]. Early studies of readability such as Gray (1935) [55] also hypothesized that vivid language could make articles easy to read. However, automatic methods to compute visual language were not feasible at that time and hence this aspect was not explored by them. Books written for student science journalists [11, 152] understandably emphasize the importance of visual descriptions.

Therefore we study visual language as one of the facets of science writing for the news. References to visual objects arise naturally in certain topics such as astronomy, plants and animals. But even for abstract topics such as research on ethics, an article's author can include visual elements, for example, a description of the attire or lab of a scientist involved in the research.

The visual property could arise due to words that refer to easily visualized elements as well as through descriptions of scenes and situations. We developed a simple measure for visual nature of a text by counting the number of visual words. Currently, the only resource of imagery ratings for words is the MRC Psycholinguistic Database [171]. It contains a list of 3,394 words each of which was rated by people for its ability to invoke an image. As such the list contains both words that have a visual nature and those that do not. With a cutoff value we adopted, of 4.5 for the Gilhooly-Logie and 350 for the Bristol Norms we obtain 1,966 visual words. So this visual word set could have low coverage for our corpus of science news. We introduce a procedure to collect a larger set of visual words from a corpus of tagged images.

Our corpus of images and their tags come from the ESP game dataset [167]. The tags were collected in a game-setting where two users individually saw the same image and had to guess words related to it. The players increased their scores when the word guessed by one player matched that of the other. This match criterion introduces some

quality control for the tags given to images, however, there is still considerable noise and non-visual words associated with the images. There are 83,904 total images and 27,466 unique tags in the corpus and the average tags per picture is 14.5. We performed filtering to find high precision image words and to also categorize them into topics.

We use Latent Dirichlet Allocation [10] to cluster the tags across all images into topics. We treat each picture as a document and the tags assigned to the picture are considered as the document's contents. We use symmetric priors set to 0.01 for both topic mixture and word distribution within each topic. We consider only images that have at least five tags. We filter out the 30 most common words in the corpus and also filter words that appear in less than four pictures. The remaining words are stemmed and clustered into 100 topics using the Stanford Topic Modeling Toolbox[26] [137].

We expected that visual words are likely to be clustered with other visual terms. So we perform filtering by rejecting or accepting the full set of words under the different topics. We use the manual annotations available with the MRC database for this purpose. We use the set of 1,966 visual words from the MRC list which we obtained using the cutoffs mentioned above. For each of the 100 topics from the topic model, we obtain the top 200 words with highest probability in that topic. We compute the precision of each topic as the proportion of the top 200 words that match the MRC list of visual words. Only those topics which had a precision of at least 25% were retained resulting in 68 visual topics. Some example topics are given in table 6.1.

Combining these 68 topics, there are 5,347 unique visual words (topics can overlap in the list of most probable words). 2,832 words from this set are not present in the MRC database. Some examples of new words in our list are 'daffodil', 'sailor', 'helmet', 'postcard', 'sticker', 'carousel', 'kayak', and 'camouflage'. For our experiments we consider the set of 5,347 words as the visual word set and also keep the information about the top 200 words in the 68 selected topics. We compute two classes of features. One is based on the total set of visual words and the other uses topic information. For a test article, we consider only the adjectives, adverbs, verbs and common nouns as candidate words for checking visual quality.

[26] http://nlp.stanford.edu/software/tmt/tmt-0.4/

landscape	**jewelry**	**shapes**
grass	silver	round
mountain	white	ball
green	diamond	circles
hill	gold	logo
blue	necklace	dots
field	chain	square
brown	ring	dot
sand	jewel	sphere
desert	wedding	glass
dirt	circle	hole
landscape	diamonds	oval
sky	jewelry	circle

Table 6.1: Sample words from three visual topics (the headings are manually assigned names)

Overall visual use. We compute the proportion of candidate words that match the visual word list as the TOTAL_VISUAL feature. We also compute the proportions based only on the first 200 words of the article (BEG_VISUAL), the last 200 words (END_VISUAL) and the middle region (MID_VISUAL) as features. We also divide the article into five equally sized bins of words where each bin captures consecutive words in the article. Within each bin we compute the proportion of visual words. We treat these values as a probability distribution and compute its entropy (ENTROPY_VISUAL). We expected these position-based features to indicate whether the placement of visual words is related to article quality.

Topic-based features. Among words from the article that we identify as visual, we also compute what proportion of the words match the list under each topic. The maximum proportion from a single topic (MAX_TOPIC_VISUAL) is a feature. We also compute a greedy cover set of topics for the visual words in the article. The topic that matches the most visual words is added first, and the next topic is selected based on the remaining unmatched words. The number of topics needed to cover 50% of the article's visual words is the TOPIC_COVER_VISUAL feature. These features indicate whether there is overwhelming presence of visual words of one kind versus combining words from different topics. Disregarding topic information, we also compute a feature NUM_PICTURES which is the number of images in the corpus where 40% of the image's tags are matched in the article. This feature is based on a similar idea as TOPIC_COVER_VISUAL but each image's tagset is considered as a (very) specific topic.

When we analyze the mean values of these features in the VERY GOOD and TYPICAL categories using the t-test procedure described in the beginning of this section, we found 8 features to vary significantly. Those that had higher mean value in the VERY GOOD category are:

Higher in VERY GOOD: BEG_VISUAL, END_VISUAL, MAX_TOPIC_VISUAL

The following features had higher mean value in the TYPICAL category.

Higher in TYPICAL: TOTAL_VISUAL, MID_VISUAL, ENTROPY_VISUAL, TOPIC_COVER_VISUAL, NUM_PICTURES

Given the studies about visual words and descriptions, we expected that the VERY GOOD articles would have more visual words. However it turns out that good writing

samples do not simply contain more visual words. They have a higher degree of visual content in the beginning and end of articles whereas typical articles have much higher visual content in the middle portions. This trend is also reflected by the entropy measure. Good articles have lower entropy for the distribution of visual words indicating that they appear in localized positions in contrast to being distributed throughout. Apart from the location-specific use of visual words, topic based features also indicate that for the VERY GOOD articles, the visual words come from only a few topics whereas TYPICAL articles show a mix of words from many topics.

6.1.2 The use of people in the story

Science writers aim to explain the impact and relevance of research findings to the readers. We hypothesized that articles that describe findings that directly affect people in some way and therefore involve explicit references and use of people in the story would be more popular. For example, the most frequent topic among our VERY GOOD samples is 'medicine and health'. Articles on this topic are often written from the view of a patient, doctor or scientist and are likely to be closer to the experiences of a reader.

Some findings from prior work also motivate the analysis of this facet. As discussed in the introduction to this chapter, the HUMAN INTEREST frame is a popular one for news reporting and its main idea is to bring the stories closer to human experiences. Studies of news frames in science journalism also point to the presence of ANECDOTAL PERSONALIZATION frame [116] where the story revolves around a person. Flesch's readability study [49] (described in Section 2.2 included a 'human interest' dimension for the same reason. Flesch computed references to people based on a list of words such as 'people' and 'folks'.

An example for a people-oriented text is below.

> Dr. Remington was born in Reedville, Va., in 1922, to Maud and P. Sheldon Remington, a school headmaster. Charles spent his boyhood chasing butterflies alongside his father, also a collector. During his graduate studies at Harvard, he founded the Lepidopterists' Society with an equally butterfly-smitten undergraduate, Harry Clench.

We approximate this facet by counting the number of explicit references to people in the test articles. We measure this value using three sources of information about an-

imacy of words. The first is named entity (NE) tags (PERSON, ORGANIZATION and LOCATION) returned by the Stanford NE recognition tool [48]. We also created a list of personal pronouns (*animate_pronouns*) such as 'he', 'myself' etc. which (almost) always indicate animate entities.

The third resource is a list containing the number of times different noun phrases (NP) were followed by each of the relative pronouns 'who', 'where' and 'which'. These counts for 664,673 noun phrases were collected by Ji and Lin (2009) [68] from the Google Ngram Corpus [88]. Ji and Lin used the data together with information about gender to identify noun phrases that refers to persons. We use a simple heuristic to obtain a list of animate (*google_animate*) and inanimate nouns (*google_inanimate*) from this list. The head of each NP is taken as a candidate noun. If the noun does not occur with 'who' in any of the noun phrases where it is the head, then it is inanimate. On the other hand, if it appears only with 'who' in all noun phrases, it is animate. Otherwise, for each NP where the noun is a head, we check whether the count of times the noun phrase appeared with 'who' is greater than each of the occurrences of 'which', 'where' and 'when' (taken individually) with that noun phrase. If the condition is satisfied for at least one noun phrase, the noun is marked as animate.

In a test article, we consider all nouns and pronouns as candidate words. If the word is a pronoun and appears in our list of *animate_pronouns*, it is assigned an 'animate' label and 'inanimate' otherwise. If the word is is a proper noun and tagged with the PERSON NE tag, we mark it as 'animate' and if it is a ORGANIZATION or LOCATION tag, the word is 'inanimate'. For common nouns, we check if it appears in the *google_animate* and *google_inanimate* lists. Any match is labelled accordingly as 'animate' and 'inanimate'. Note that this procedure may leave some nouns without any labels.

Our features are counts of animate tokens (ANIM), inanimate tokens (INAMIN) and both these counts normalized by total words in the article (ANIM_PROP, INANIM_PROP). Three of these features had significantly higher mean values in the TYPICAL category of articles: ANIM, ANIM_PROP and INANIM_PROP. We found upon observation that several articles that talk about government policies involve a lot of references to people but are often in the TYPICAL category. These findings suggest that the 'human' dimension might

need to be computed not only based on references to people but also based on other words that are commonly associated with people's experiences.

6.1.3 Beautiful language

Beautiful phrasing and word choice can entertain a reader and leave a positive impression. These aspects are separate categories in the Six Traits rubric showing that there is great emphasis by the raters on elegant language use. For example, the snippet below can be said to be creatively written.

> When I was in the sixth grade – could it really have been 30 years ago? – I swirled around the playground in an oversized crocheted number. Worn casually with my elephant bellbottoms, a flower-power T-shirt and waist-length hair, the poncho struck a subtle but powerful blow against an older, more conservative generation of jacket wearers. Sleeves? So uptight. The poncho is no longer a rebel.

However, detecting such writing could be quite difficult and subjective. Different linguistic realizations could contribute to a perception of beautiful writing and be hard to separate out. We implement a method based on a simple idea that creative words and phrases are sometimes those that are used in unusual contexts and combinations or those that sound unusual.

We compute measures of unusual language both at the level of individual words and for the combination of words in a syntactic relation.

Word level measures: Unusual words in an article are likely to be those with low frequencies in a background corpus. We use the full set of articles (not only science) from year 1996 in the NYT corpus as a background (these do not overlap with our corpus for article quality). We also explore patterns of letters and phoneme sequences with the idea that unusual combination of characters and phonemes could create interesting words. We used the CMU pronunciation dictionary [168] to get the phoneme information for words and built a 4-gram model of phonemes on the background corpus. Laplace smoothing is used to compute probabilities from the model. However, the CMU dictionary does not contain phoneme information for several words in our corpus. So we also compute an

Low frequency	High perplexity-phonemes	High perplexity-letters
undersheriff	showroom	kudzu
woggle	yahoo	muumuu
ahmok	dossier	qipao
hofman	powwow	yugoslav
volga	plowshare	kohlrabi
oceanaut	oomph	iraqi
trachoma	chihuahua	yaqui
baneful	ionosphere	yakuza
truffler	boudoir	jujitsu
lacrimal	superb	oeuvre
corvair	zaire	yaohan
entomopter	oeuvre	kaffiyeh

Table 6.2: Top rated unusual words according to our three measures

approximate model using the letters in the words and obtain another 4-gram model.[27] Only words that are longer than 4 characters are used in both models and we filter out proper names, named entities and numbers.

During development, we analyzed the articles from an entire year of NYT, 1997, with the three models to identify unusual words. Table 6.2 lists the words with lowest frequency and those with highest perplexity under the phoneme and letter models.

For computing the features, we consider only nouns, verbs, adjectives and adverbs. We also require that the words are at least 5 letters long and do not contain a hyphen[28]. Three types of scores are computed. FREQ_NYT is the average of word frequencies computed from the background corpus. The second set of features are based on the phoneme model. We compute the average perplexity of words under the model, AVR_PHONEME_PERP_ALL. In addition, we also order the words in an article based on decreasing perplexity values and the average perplexity of the top 10, 20 and 30 words in this list are added as features (AVR_PHONEME_PERP_10, 20, 30). We obtain similar features from the letter n-gram model (AVR_CHAR_PERP_ALL, AVR_CHAR_PERP_10, 20, 30). In

[27]We found that higher order n-grams provided better predictions of unusual nature during development.

[28]We noticed that in this genre several new words are created using hyphen to concatenate common words.

phoneme features, we ignore words that do not have an entry in the CMU dictionary.

Word pair measures: Next we attempt to detect unusual combinations of words. We do this calculation only for certain types of syntactic relations—a) nouns and their adjective modifiers, b) verbs with adverb modifiers, c) adjacent nouns in a noun phrase and d) verb and subject pairs. Counts for co-occurrence again come from NYT 1996 articles. The syntactic relations are obtained using the constituency and dependency parses from the Stanford parser [33, 75]. To avoid the influence of proper names and named entities, we replace them with tags (NNP for proper names and PERSON, ORGANIZATION, LOCATION for named entities). The named entities were identified using the Stanford named entity recognition tool.

We treat the words for which the dependency holds as a (auxiliary word, main word) pair. For adjective-noun and adverb-verb pairs, the auxiliary is the adjective or adverb; for noun-noun pairs, it is the first noun; and for verb-subject pairs, the auxiliary is the subject. Our idea is to compute usualness scores based on frequency with which a particular pair of words appears in the background.

Specifically, we compute the conditional probability of the auxiliary word given the main word as the score for likelihood of observing the pair. We consider the main word as related to the article topic, so we use the conditional probability of auxiliary given main word and not the other way around. However, the conditional probability has no information about the frequency of the auxiliary word. So we apply ideas from interpolation smoothing [18] and compute the conditional probability as an interpolated quantity together with the unigram probability of the auxiliary word.

$$p(aux|main) = \lambda * p(aux|main) + (1 - \lambda) * p(aux)$$

The unigram and conditional probabilities are also smoothed using Laplace method. We tune the λ value (a separate one for each type of word pair) to optimize data likelihood using the Baum Welch algorithm and use the pairs from NYT 1997 year articles as a development set. The λ values across all types of pairs tended to be lower than 0.5 giving higher weight to the unigram probability of the auxiliary word.

ADJ-NOUN	ADV-VERB	NOUN-NOUN
hypoactive NNP	suburbs said	specification today
plasticky woman	integral was	auditory system
psychogenic problems	collective do	pal programs
yoplait television	physiologically do	steganography programs
subminimal level	amuck run	wastewater system
ehatchery investment	illegitimately put	autism conference
multistage process	straighter make	timbre changes
aquacultural products	secret talk	pulmonology department
caplike form	holy keep	monkeypox case
apomorphine treatment	cerebrally felt	monkeypox cases
antispam operations	norepinephrine knew	strontium levels

SUBJ-VERB		
blog said	briefer said	hr said
knucklehead said	lymphedema have	permissions have
steganography have	monkeypox had	ipso is
neuroscientist said	cybertrainer make	

Table 6.3: Unusual word-pairs from different categories

Based on our observations on the development set, we picked a cutoff of 0.0001 on the probability (0.001 for adverb-verb pairs) and consider phrases with probability below this value as unusual. For each test article, we compute the number of unusual phrases (total for all categories) as a feature (SURP) and also this value normalized by total number of word tokens in the article (SURP_WD) and normalized by number of phrases (SURP_PH). We also compute features for individual pair types and in each case, the number of unusual phrases is normalized by the total words in the article (SURP_ADJ_NOUN, SURP_ADV_VERB, SURP_NOUN_NOUN, SURP_SUBJ_VERB).

A list of the top unusual words under the different pair types are shown in Table 6.3. These lists were computed on pairs from a random set of articles from our corpus. Several of the top pairs involve hyphenated words which are unusual by themselves, so we only show in the table the top words without hyphens.

All these features are different between the two categories as expected.

Higher in VERY GOOD: AVR_PHONEME_PERP_ALL, AVR_CHAR_PERP_(ALL, 10), SURP, SURP_PH, SURP_WD, SURP_ADJ_NOUN, SURP_NOUN_NOUN, SURP_SUBJ_VERB

Higher in TYPICAL: FREQ_NYT

The average perplexity of words from the VERY GOOD articles is higher under both the character and the phoneme models. The average frequency of these words in the background corpus is also lower. For the word pair based features, the proportion of unusual phrases is also higher in the VERY GOOD articles. These findings indicate that unusual word phrases as hypothesized are associated with the good samples in our corpus.

6.1.4 Sub-genre

This aspect differentiates articles at the organization level and abstracts away from individual words and sentences. There are several sub-genres in science writing [152]: short descriptions of discoveries, longer explanatory articles, narratives, stories about scientists, reports on meetings, review articles and blog posts. We expected that some of these sub-genres could be more appealing to readers. For example, a narrative may be more interesting to a reader as he can involve himself with the story line and characters. A snippet from a narrative article in our science journalism corpus is shown below.

> Mr. Jousse became one of the world's foremost urban lighting experts by accident. A native of Paris, he landed a job in 1963 with the city's engineering division after graduating from college, helping widen and deepen the city's canals. He later had jobs supervising 3,000 garbage collectors and creating pedestrian streets. In 1981, a supervisor asked him to change course once again.

There are several studies on genre prediction amd mostly on the news domain [72, 121]. These methods use part of speech, pronouns and stop words as features. Rather than include features that are related to genre differences we choose to directly compute scores for some genres of interest in our corpus. We compute simple measures to indicate three genres—narrative, attribution and interview.

Narrative texts typically have characters and events [111]. Based on this idea, we compute a score for the narrative nature of a text based on two factors—entities (pronouns

and proper names) and past tense. We count the number of sentences where the first verb in surface order is in the past tense. Then among these sentences, we pick those which have either a personal pronoun or a proper noun before the target verb (again in surface order). The proportion of such sentences in the text is taken as the score (named NARRATIVE).

We also developed a measure to identify the degree to which the article's content is attributed to external sources compared to the author's own statements. Attribution to other sources is frequent in the news domain since many comments and opinions are not the views of the journalist. As we already discussed, the RESPONSIBILITY frame which is related to attribution is rather common in news reporting [145]. For science journalism, attribution becomes even more important since the research findings were obtained by scientists and reported in a secondhand manner by the journalists. So we compute a score (ATTRIB) to indicate the level to which the author talks directly about the subject compared to using attributive statements from the scientists. This score is the proportion of sentences in the article that have a quotation mark, or the words 'said' and 'says'.

We also compute a score to indicate if the article is the account of an interview. There are easy clues in NYT for this genre with paragraphs in the interview portion of the article beginning with either 'Q.' (question) or 'A.' (answer). We count the total number of 'Q.' and 'A.' prefixes combined and divide the value by the total number of sentences (INTERVIEW). When either the number of 'Q.' tags is zero or 'A.' tags is zero, the score is set to zero.

All three scores are significantly higher for the TYPICAL class.

6.1.5 Affective content

The writing in an article can also evoke emotions and sentiment in a reader. For example, articles detailing research on health, crime, ethics and well-being can involve and discuss issues that have a lot of sentiment value and be more appealing to a reader. A snippet with high sentiment value is shown below.

> "Although it could be argued that there is little to lose in this tragic situation," he wrote, "my personal view is that there is a significant risk of causing pain or dis-

> tress if the treatment is given and very little prospect of any benefit." Medicine is a constant trade-off, a struggle to cure the disease without killing the patient first. Chemotherapy, for example, involves purposely poisoning someone – but with the expectation that the short-term injury will be outweighed by the eventual benefits.

We compute features for sentiment value using three lexicons. Two of these, MPQA [172] and General Inquirer [153] give lists of positive and negative sentiment words. The third resource is a set of words associated with emotions and were obtained from FrameNet (Emotion frame) [5]. The sizes of these lexicon are 8221, 5395, and 653 words respectively. We compute the counts of positive, negative, polar, and emotion words, each normalized by the total number of content words in the article (POS_PROP, NEG_PROP, POLAR_PROP, EMOT_PROP). We also include the proportion of emotion and polar words taken together (POLAR_EMOT_PROP) and the ratio between count of positive and negative words (POS_BY_NEG) as features.

The significant features are listed below:

Higher in VERY GOOD: NEG_PROP, POLAR_PROP, EMOT_POLAR_PROP

Higher in TYPICAL: POS_BY_NEG, EMOT_PROP

VERY GOOD articles do turn out to have more sentiment words. It should also be noticed that the proportion of positive words does not vary between categories but the VERY GOOD articles have higher proportions of negative sentiment words. A similar trend is the TYPICAL articles having higher values for positive to negative word ratio. However, emotion words are more frequent in the TYPICAL articles.

6.1.6 Amount of research content

Science news cannot convey the full depth of research done on a topic in the way that academic publications do. For a lay audience, a science writer chooses the most relevant findings and methods of the research to include in the article and also interleaves the research information with details about the relevance of the finding, people involved in the research and general information about the topic. So the degree of explicit research descriptions in the articles varies considerably.

> The study is being published in the April issue of the journal Psychological Science. The findings seem to fall in line with the idea that dreams express complicated desires and unfulfilled wishes, as Freud, who called dreams the "royal road to the unconscious," noted long ago. But Dr. Wegner does not completely agree with that assertion.

To test how this aspect is related to quality, we count references to research methods and research people in the article. We use the research dictionary that we introduced during corpus creation (Chapter 3) as the source of research-related words. We count the total number of words in the article that match the dictionary (RES_TOTAL) and also the number of unique matching words (RES_UNIQ). We also normalize these counts by the total words in the article and create features RES_TOTAL_PROP and RES_UNIQ_PROP.

All four features have significantly higher values in the VERY GOOD articles which indicate that popular articles are also associated with a great amount of direct research content and explanations.

6.2 Validating the features

As we noted in the introduction to this chapter, we have deliberately only used features which we hope we can relate to quality in a direct manner. Having a large class of features where individual ones do not have a clear relationship to a writing facet will limit our ability to claim if any definable facet is indicative of text quality. Rather the analysis will only denote individual myopic features as significantly predictive. For example, suppose that we find personal pronouns to occur significantly more often in VERY GOOD versus the TYPICAL category of articles. This result does not necessarily indicate that a narrative style is indicative of good quality or that references to people are more common in good samples. However, for tasks such as text quality prediction, such interpretable results are preferrable. In this section, we describe an annotation study where we directly studied if our features are capturing the intended aspect with good accuracy. During this annotation, our aim is to only understand the representative nature of the features separate from whether the feature is indicative of text quality.

For this analysis, we selected eight features, one from each of our six facets, with the exception of 'beautiful language' and 'affective content'. For beautiful language, we select two features: AVR_CHAR_PERP_ALL which indicates the average perplexity of words under the ngram character model and SURP_WD which is a word-pair related feature which measures the number of unusual phrases (normalized by number of words). For affective content, we select the features measuring the total proportion of polarity words (POLAR_PROP) and the proportion of total words which have negative sentiment (NEG_PROP).

To obtain text examples, we selected a random sample of articles from our corpus (without regard to quality categories). However, we biased the sample to be representative of different topics in our corpus. We utilize the set of "science" tags from Chapter 3 (Section 3.1.2) for this purpose. These tags are taken from the NYT corpus metadata and indicate a minimal set of science related topics in the NYT. There were 14 tags in that set. We exclude the 'Research' tag since it is does not indicate a specific topic. For each of the remaining tags, we randomly sample 25 articles from the corpus which contain that tag. In this way, we obtain a representative small sample of our corpus with a total of 325 articles.

Since it would be difficult to judge the presence of a facet in a full article or further to indicate its extent in the article, we create smaller snippets from the articles, each of size 200 words. We create snippets starting from each paragraph boundary in the article and do not truncate the snippet in the middle of a sentence. The resulting snippets are quite coherent and a total of 6192 snippets were obtained.

For each feature, we compute its value for all the snippets. Then, we select the 50 snippets with highest feature value, the 50 with lowest value for the feature and 50 samples randomly chosen without regard to feature value.[29] We provided these snippets in random order and asked annotators to indicate the degree to which the facet represented by the feature is present in the snippet. For example, for the affective content feature, we asked an annotator to rate the passage for the degree to which sentiment and emotion is present in the snippet. The annotators used a scale from 1 to 10 where 10 indicates that the facet is present to a very high degree and 1 indicates that the facet is almost absent.

[29]We select only one snippet per article to avoid having the annotation data biased towards a few articles only. The next high ranking snippet from a different article than those already selected is chosen.

Note that our annotation procedure is based on texts ranked high and low according to certain feature values. An alternative method is to first directly obtain ratings for each facet on a collection of snippets and then compute the extent to which our features reflect these ratings. In the latter approach, it is unclear how large a collection we should annotate in order to obtain samples which have high and low degree of presence for all the aspects that we consider. So we choose our two step approach of first obtaining feature values on the texts and then estimating the accuracy of the induced rankings.

Our annotators were undergraduate students from University of Pennsylvania's engineering and pschology departments and are all native speakers of English. During a training phase, each student was assigned two aspects which they studied in detail. A description of the facet was provided together with example snippets that were manually chosen to reflect high, low and medium presence of the facet. Each facet was also assigned to two different annotators. They annotated a sample of 10 snippets individually and the two annotators who rated the same facet discussed their ratings with each other.[30] Even during the training sessions, we found that the annotators had reasonable agreement in their ratings and were able to discuss to resolve differences.

After training, each annotator annotated the 150 snippets belonging to top, bottom and random values (each 50) of a feature. Another annotator annotated a random sample of 30 snippets (from the 150) in order to measure agreement. If a feature captures a particular aspect then the snippets ranked at the top should receive higher ratings from annotators compared to those ranked by the feature as low. We include the set of random snippets to check the prevalence of an aspect. If any snippet chosen at random has a high value for the aspect from the annotators, it would indicate that the aspect is highly prevalent in the texts in our corpus. So a feature based on this aspect is unlikely to be useful for differentiating the articles.

The results are shown in Table 6.4 for the eight selected features. The second column indicates annotator agreement which we measure as the Pearson correlation between the ratings of the two annotators on the common 30 snippets. A '$*$' indicates that the correlation was significant with p-value less than 0.05. The next three columns indicate

[30]These snippets were chosen from a different set of articles than those used for final annotation.

Feature	Agree-ment	Mean ratings from annotator			Significance
		Top (T)	Bottom (B)	Random (R)	
TOTAL_VISUAL	0.57*	4.72	1.88	2.84	T > B, T > R, B < R
ANIMATE_PROP	0.94*	6.72	1.30	4.04	T > B, T > R, B < R
NARRATIVE	0.78*	7.34	3.72	4.52	T > B, T > R
AVR_CHAR_PERP_ALL	0.09	4.50	4.62	4.30	
SURP_WD	0.47*	4.80	4.08	4.12	T > B, T > R
POLAR_PROP	0.71*	4.68	1.96	2.86	T > B, T > R, B < R
NEG_PROP	0.69*	4.96	1.28	2.48	T > B, T > R, B < R
RES_TOTAL_PROP	0.71*	3.84	1.30	2.46	T > B, T > R, B < R

Table 6.4: Agreement (Pearson correlation) of annotators and mean values of ratings for the different splits in feature value. The last column indicates whether the ratings for the splits are significantly different. Significant correlations in the second column are marked with a '*'

the mean value of the annotator rating for the top, bottom and random snippets. The last column indicates whether the mean value for *top* ranked snippets is significantly higher than *bottom* ranked snippets ($T > B$) and if the *top* and *bottom* snippets have ratings significantly different from *randomly* chosen snippets. High or low trends are indicated by $>$ and $<$ symbols. The values in two classes of snippets were compared using a two-sided t-test and a p-value of less than 0.05 was taken to indicate significance.

We find that for most of our features, the two annotators had high agreement in their judgements of whether the text ranked high or low with regard to the corresponding facet. Most of these correlations between the annotators' ratings are 0.5 and above. The highest agreement is for animacy feature reaching 0.9 correlation. For the AVR_CHAR_PERP_ALL feature, there is no correlation at all between the annotators. The proportion of visual words and the SURPH_WD features have around 0.5 correlation. Narrative sub-genre, polarity and research content features have 0.7 correlation.

For the differences between top, bottom and random snippets, most of the features showed the desirable trends. The annotators rated the top ranked snippets according to feature value as having high presence of the aspect and the bottom snippets as having

much lower presence of the aspect. Similarly, both top and bottom snippets are rated significantly different from random snippets indicating that these features create useful distinction between texts according to the facet they represent. The only feature where no significant results were obtained is the one for unusual words. Note also that annotators did not have any agreement for ratings for this feature. This result indicates that either this feature does not capture the 'unusual words' aspect or that people do not perceive unusual words as related to beautiful writing. Notably, all the features with the exception of 'beautiful writing' are designed to reflect a facet of writing (such as sentiment) without reference to whether the text is considered as interesting or of high/low quality. However in the case of 'beautiful language' features, we are directly asking annotators to judge the attractive nature of the writing and this could increase the variability in ratings according to a person's preferences and opinions. Future work should focus on how different aspects can be annotated separate from questions of quality judgement.

6.3 Experimental setup

We perform two types of classification tasks for text quality. We briefly review the division of our corpus into development and test sets which we outlined in Chapter 3.

Any topic: Here the goal is to separate out VERY GOOD versus TYPICAL articles without regard to topic. The test set contains the 4,153 VERY GOOD (or GREAT) articles and we randomly sample 4,153 articles from the TYPICAL category to comprise the negative set.

Same topic: Here we use the topic-paired VERY GOOD and TYPICAL articles. The goal is to predict which article in the pair is the VERY GOOD article. For the test set, we selected 41,530 pairs.

Development data: We randomly selected 100 VERY GOOD articles and their paired (10 each) TYPICAL articles from the topic-normalized corpus. Overall, these constitute 1000 pairs which we use for developing the same-topic classifier. From these selected pairs we take the 100 VERY GOOD articles and sample 100 unique articles from the TYPICAL articles making up the pairs. These 200 articles are used for tuning the any-topic classifier.

Feature set	Any-topic	Same-topic
interest-science	77.5	70.2
Ablation tests		
− visual nature	77.3	68.1
− use of people	77.1	69.9
− beautiful language	73.4	64.9
− sub-genre	75.8	68.2
− affective content	76.2	67.8
− research	72.7	68.4

Table 6.5: Accuracy of the interest features (interest-science) and ablation tests for different subsets of features ('-' indicates that the feature set was removed from the interest-science features)

6.4 Interest measures and text quality

This section reports the results of classification experiments using the interest features we introduced above. The baseline random accuracy for both our tasks is 50%.

We use a SVM classifier with a radial basis kernel (implementation in R [132]) for our experiments. The regularization and kernel parameters were tuned using cross validation on the development data.

6.4.1 Accuracy on the two tasks

Table 6.5 gives the 10-fold classification results on the test sets using the chosen best parameters. The set of all features (related to the six facets) that we described above are named as the *interest-science* category. We also report the results from ablation tests to understand which classes of features greatly impact classification performance.

The *interest-science* features give remarkable performance. For the 'any-topic' setup, we obtain accuracies 27% above the baseline and for the topic normalized corpus, the improvement is 20%.

The ablation tests indicate that overall when the *beautiful language* features are re-

moved, the performance decreases the most (4% for 'any-topic' setup and 5% for the 'same-topic' task). For the 'any-topic' setup, the degree of explicit research discussion is also highly impactful. Many of the other features also lead to lower accuracies when removed from the classifier. The one exception is the feature related to use of people, which although could be annotated with high agreement, was not useful for making the distinction between our categories.

6.4.2 Accuracy and topic similarity

The topic normalized corpus contains article pairs with varying similarity between them. In this section, we investigate the relationship between topic similarity and accuracy of prediction in detail. In an information retrieval setting, a system would need to rank articles that are similar in topic. We use this experiment to understand how the performance of text quality prediction changes as the articles get more and more similar in content and topic.

We create ranges of similarity values and collect pairs with similarity within each range into a corresponding bin. We compute the 10-fold cross validation predictions using the different feature classes and the overall *interest-science* feature set and collect the predicted values across all the folds. Then we compute accuracy of examples within each bin based on the predicted values. These results are plotted in Figure 6.1. *int-science* refers to the full set of features and the results from the six feature classes are also indicated.

As the similarity increases, the prediction task becomes harder. Using all the features, the accuracy around 70% for pairs above 0.3 similarity and 81% when the similarity is about 0.05 to 1.0.

Most individual feature classes also have lower performance with increasing similarity with the exception of three—affective content, visual nature and research descriptions. For these features, the accuracies improve with higher similarity; sentiment features give 42% accuracy for pairs with similarity 0.05-0.1 and 58% for pairs above 0.4 similarity, accuracy of research features goes from 40% to 61% for the same similarity values. Visual features increase in accuracy from 48% to 64%. These three sets of features appear to have special benefits for ranking articles during information retrieval.

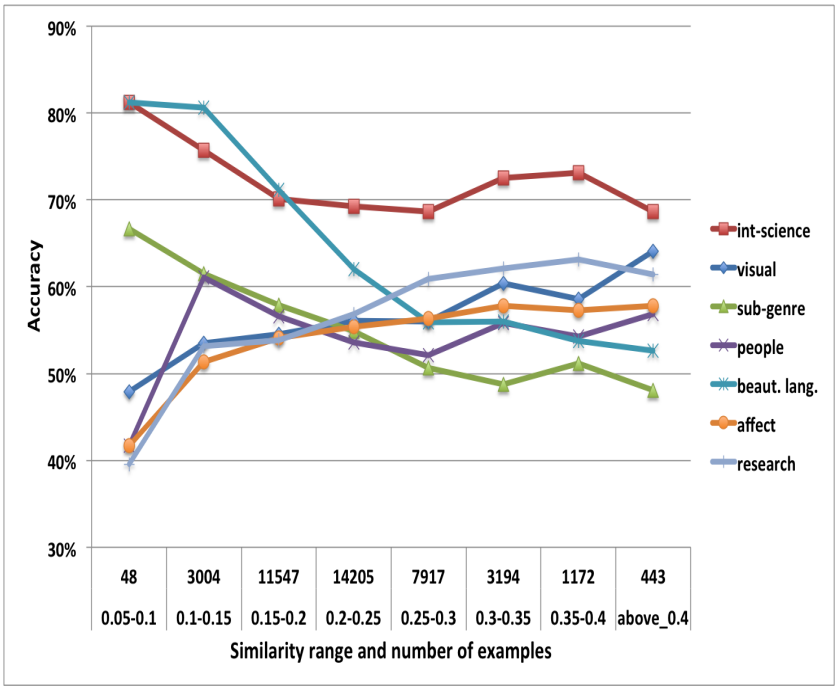

Figure 6.1: Accuracy of feature classes on pairs with different similarity

6.5 Comparing and combining our features with prior work

We now report experiments where we compare the accuracy of the features that we have developed with other methods proposed in prior work for predicting different aspects of quality. We detail these features and also explain how they vary between our VERY GOOD and TYPICAL categories using the same set of articles and t-test procedure which we used in Section 6.1 when we introduced the *interest-science* features. Then we present classification experiments to show the strength of these features from prior work and the accuracy of combining our interest features with these. We do this analysis in two parts. In Section 6.5.1, we explore features introduced for other writing aspects such as readability and well-written nature and indicators of interesting fiction articles. Separately, we examine the influence of the topic of the article upon reader interest in Section 6.5.2

A noteworthy point in these experiments is that the features are taken from studies of different writing quality aspects in prior work. However, they are not trained on the gold standard data for that aspect. For example, we do not train the readability features on data with educational grade levels. Rather the features are trained on our science

corpus and based on our categories as the gold standard. Therefore while a feature was proposed for readability or well-written nature, it may capture a different distinction in our corpus of science news. Sometimes, this setup may lead to difficulty in interpreting a feature, because it has an opposite trend than expected from prior work. However, we believe that training these features on our corpus provides the best amount of training data.

6.5.1 Features for readable, well-written and interesting texts

Readability (16 features). As we described in Chapter 2, there are numerous studies on readability prediction. We computed three types of measures as a representative set for comparison. The first, lexical type features include number of tokens, type-token ratio, average word length and average and maximum sentence length (TOKENS, TTRATIO, WLEN, AVG_SLEN, MAX_SLEN). These counts are components of most readability formulae. We also add language model likelihoods to capture word familiarity. We use two language models trained on Wall Street Journal and the Associated Press as in Pitler and Nenkova (2008) [124]. The unigram likelihoods from these models are features: LANG_WSJ and LANG_AP. The second class contains syntactic measures developed by Schwarm and Ostendorf (2005) [144] to indicate sentence complexity. These are average parse tree height (PARSE_HT), average number of noun phrases (AVG_NP) and verb phrases (AVG_VP) per sentence and also average number of subordinate clauses (AVG_SUB). The third class of features are related to text cohesion. It comprises average overlap between adjacent sentences based on words computed in three ways: number words in common (OVERLAP_WD_COUNT) and as cosine similarity between word counts (OVERLAP_WD_COSINE) and as number of common nouns and pronouns (OVERLAP_ENTITIES). These features also include the average number of definite articles per sentence (AVG_DEF) and number of pronouns per sentence (AVG_PRP).

The features which vary significantly between the VERY GOOD and TYPICAL articles are:

Higher in VERY GOOD: AVG_SLEN, AVG_SUB, AVG_VP, PARSE_HT, OVERLAP_WD_COUNT, OVERLAP_WD_COSINE, OVERLAP_ENTITIES, AVG_DEF, AVG_PRP

Higher in TYPICAL: TTRATIO

Some features have the expected trends as per readability, for example, the VERY GOOD articles have greater word overlap between adjacent sentences. The TYPICAL articles have a higher type token ratio. But we also find that factors that readability work considers as associated with greater reading difficulty, higher frequency of subordinate clauses, complex parse structure and definite articles are more frequent in the VERY GOOD samples.

Well-written nature (23 features). Here, the idea is to predict texts of good and acceptable writing in contrast to "easy to read" versus difficult distinction. We use two classes of features proposed for this aspect, both related to discourse phenomenon. The first set comes from the Entity Grid model introduced by Barzilay and Lapata (2008) [7]. The probabilities of different types of entity transitions are taken as features. This set has 16 features which we refer to as ENTITY_GRID. The entity grid itself was created for our articles using the Brown Coherence Toolkit [42]. The other class of features are discourse relation likelihoods and counts introduced by Pitler and Nenkova (2008) [124]. In that work, these features were computed using gold standard discourse annotations from the Penn Discourse Treebank [128]. Both implicit and explicit discourse relations are annotated in this corpus and both were used to compute features. For our corpus, we identify the explicit relations using the *addDiscourse*[31] [125] tool and use only these explicit relations for feature computation. Our features are total relations per sentence (DISC_RELS_PROP) and similarly for individual relations, Expansions (EXPN_RELS_PROP), Contingencies (CONT_RELS_PROP), Temporal (TEMP_RELS_PROP) and Comparisons (COMP_RELS_PROP). These are the four main classes of discourse relations in the PDTB. We also include the unigram (UNI_DISC_LIK) and multinomial likelihood (MULT_DISC_LIK) of the relations in the test article based on a language model trained on the explicit relations from the PDTB. Detailed descriptions of these features can be found in Chapter 2.

The t-tests show that the VERY GOOD articles have more discourse relations and greater likelihood under the discourse relations language model. Most of the entity grid transitions were also more frequent in this category. These findings are consistent with results reported by Pitler and Nenkova (2008) [124] who study both the discourse and entity grid features in their paper.

[31] http://www.cis.upenn.edu/~nlp/software/discourse.html

Higher in VERY GOOD: All ENTITY_GRID transitions except '−−', DISC_RELS_PROP, CONT_RELS_PROP, TEMP_RELS_PROP, MULT_DISC_LIK

Higher in TYPICAL: ENTITY_GRID '−−' transition

Interesting fiction (22 features). We include the features used by McIntyre and Lapata (2009) [104] for predicting interest ratings on fiction articles (short fairy tales). They include counts of different syntactic items and relations, and token categories from the MRC psycholinguistic database. In that work, the features at article level were created by summing up the counts across all tokens. However, our articles are of widely varying sizes. We therefore use the average values of the token level scores rather than the sum.

McIntyre and Lapata found that most of their 23 features correlate positively with the interest ratings given by people. However the trends were rather different for our corpus.

Higher in VERY GOOD: adjective tokens, adverb tokens and types, frequency from Brown corpus, meaningfulness score 2[32]

Higher in TYPICAL: noun tokens and types, verb types, number of objects, number of Brown categories, concreteness, imagery, meaningfulness score 1

Only a few features distinguish between our categories and the trends are different from that observed on fiction articles. For example, McIntyre and Lapata find that adjective and adverbs counts were among the few features not correlated with interest, and noun, imagery and concreteness features were positively related.

Classification results

We present classification results using the features introduced above as individual classes and in combination with the interest features which we introduced in this chapter. Table 6.6 shows these results.

For readability, well-written and interesting fiction we trained individual SVM based classifiers and also create SVM models for their combination with the *interest-science* features. For each model, the parameters were tuned on the development set.

The readability, well-written nature and interesting fiction classes provide good accuracies 61% and above. Combinations of these three feature sets improves performance

[32]Two types of scores were computed.

Feature set	Any-topic	Same-topic
Interest-science	77.5	70.2
Readable	66.8	64.1
Well-written	61.4	62.3
Interest-fiction	68.8	64.6
Readable + well-written	65.8	67.2
Readable + well-written + Interest-fiction	74.2	72.0
Readable + well-written + Interest-science	76.4	76.3
All writing aspects	76.0	78.1

Table 6.6: Accuracy of interest features versus those developed for other aspects of quality

compared with individual classes giving over 70% accuracy for both setups—any-topic and same-topic. The genre-specific *interest-science* features are individually much stronger than the other classes. Particularly for the any-topic setup, the accuracy is 77%, 3% better than the combination of readability, well-written and interesting fiction features. When all four dimensions are combined (the marked "*All writing aspects*" in Table 6.6), the accuracy is even better with 76% accuracy for the any-topic task and 78% for the topic paired task. Note that for the any-topic setup, the *interest-science* features are individually better than when features from all four classes used together.

These results provide evidence that reader interest related features are individually most successful for differentiating the article categories in our corpus. We would expect that interest related features capture a different set of quality aspects compared to well-written and organized nature. This hypothesis was also confirmed in our results. In fact, combining all classes of features leads to the best performance in quality prediction.

Which features are most useful?

We analyzed our results in the classes above to provide a view of how features proposed for different aspects of quality perform for our task. However, these divisions are not rigid and features from different classes obviously interact. So we also attempted to find a subset of features from the full class which provides very high accuracy. To do this

analysis, we first ranked the features by their fscores computed on the development set. Fscores are frequently used for feature selection for SVM classifiers [19] and the value for a feature i is computed as:

$$\text{fscore}(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \tag{6.1}$$

$\bar{x}_i$ is the average value of the feature across all examples. $\bar{x}_i^{(+)}$ and $\bar{x}_i^{(-)}$ are the average feature values in the positive and negative examples respectively. $x_{k,i}^{(+)}$ is the value of the feature on the k^{th} positive example. The fscore numerator represents the variation in the feature's values between the positive and negative classes. The denominator indicates how much variability is there in the feature's values within each class. Higher fscores indicate features with greater discriminatory power.

One drawback of fscores is that they do not consider the relationship between features and compute importance based on individual features only. However, fscore-based feature selection works rather well in practice. We computed fscores for each feature on the development set and the top 15 features for the two classification setups are shown in Table 6.7. We have listed them in groups ignoring the actual order so that the feature names are easier to interpret. The beautiful language features were at the top of the list for both tasks.

Most of the features in the top lists are from the *interest-science* class and the list of features overlap considerably for the two tasks. Two subclasses of interest-science features are prominent in this list—research content and beautiful language. Interest-fiction features are also among the top features for both tasks. One entity grid transition is also in the top list for each task.

Then we trained classifiers with increasing number of features selected in order of these importance scores. We add features in bins of size 5, the first classifier has 5 features, second has 10 and so on. For each classifier, the parameters are tuned on the development set. Figure 6.2 shows the performance of these classifiers for our two tasks.

For both setups, we reach the highest accuracy when all the features are included. There are a total of 102 features across all the sets. For the first 50 of the features, we receive considerable accuracy improvement during classification. At this point, the accu-

Any-topic setup	Same-topic setup	Any-topic setup	Same-topic setup
Beautiful language features		**Interesting fiction features**	
SURP_PH	SURP_PH	no. of syllables	no. of syllables
SURP_WD	FREQ_NYT	no. of phonemes	no. of phonemes
FREQ_NYT	SURP_WD	no. of objects	
SURP_SUBJ_VERB	SURP_SUBJ_VERB		
SURP_NOUN_NOUN	SURP_NOUN_NOUN	**Research degree features**	
AVG_CHAR_PERP_10	AVG_CHAR_PERP_20	RES_UNIQ	RES_UNIQ
AVG_CHAR_PERP_20	AVG_CHAR_PERP_10	RES_TOTAL	RES_TOTAL
AVG_PHONEME_PERP_10	AVG_CHAR_PERP_30		
AVG_CHAR_PERP_30	AVG_PHONEME_PERP_10	**Entity grid features**	
	SURP	OO transition	$-O$ transition

Table 6.7: Top 15 features by fscore (grouped into feature classes). In the entity grid, OO indicates a transition from object role in previous sentence to object role in current sentence and $-O$ indicates the entity is absent in previous sentence and is an object in current sentence

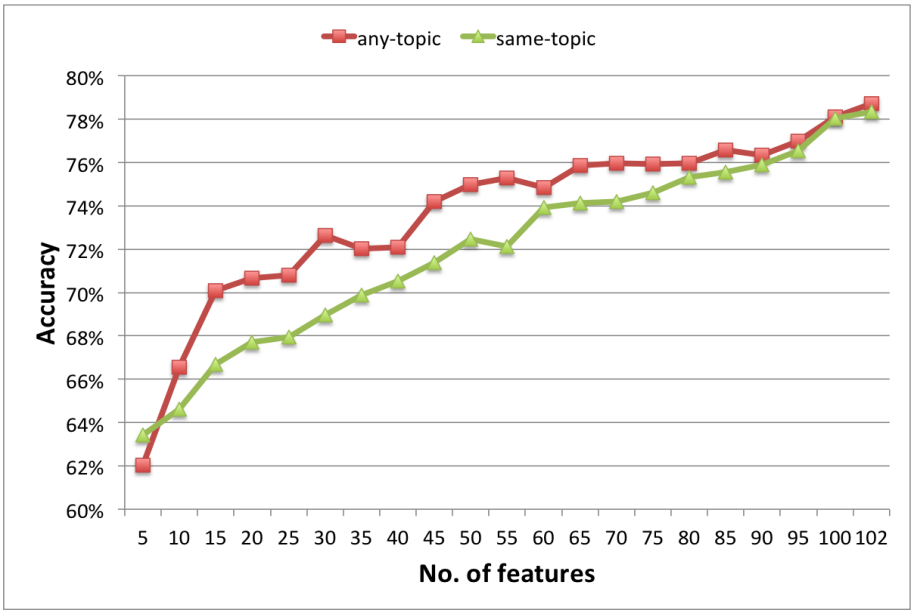

Figure 6.2: Accuracy with increasing number of important features

feature set	No. of features	No. of features in top list	
		any-topic	same-topic
well-written	23	12 (52.2%)	14 (60.9%)
interest-science	41	24 (58.5%)	23 (56.1%)
interest-fiction	22	9 (40.9%)	7 (31.8%)
readability	16	5 (31.0%)	6 (37.5%)

Table 6.8: Features from different classes that are in the top 50 list for the two classification tasks

racy is 75% for any-topic setup and 72% for same-topic. The addition of the remaining 52 features only improves the accuracy by a few points, 3% for any-topic and 6% for same-topic. So we provide a detailed split of these top 50 features according to the classes they were taken from. These results are shown in Table 6.8.

The top 50 lists includes features from all the aspects of quality but some have stronger presence compared to others. For the well-written and interesting-science classes, more than half their features are selected in the top list. The highest membership for an individual class is well-written where 60.9% of its features are chosen for the same-topic setup. The interest-fiction and readability features have lower membership, 30-40% of these features are in the top list. One reason could be that readability is designed to separate texts for different audience capacities while all the texts in our corpus are aimed at the same level of readers. The interest-fiction class is a mix of different token level scores. Some of the distinctions such imagery and concreteness scores which they compute using the MRC lexicons could have low coverage on our corpus leading to lower fscores for these features.

However, we find that accuracies improve continuously with addition of features and the highest accuracy is obtained when all the features are included. This result shows that whether we want to select an interesting article from all topics of the newspaper or rank articles within the same topic topic during information retrieval, we would require features from all these quality dimensions to make an accurate prediction. This finding conforms with one of the main themes of this thesis that different aspects of quality

Tag	Articles	Tag	Articles
Medicine and Health	22	Computers and the Internet	4
Research	18	Doctors	4
Space	14	Drugs (Pharmaceuticals)	4
Science and Technology	13	Evolution	4
Physics	10	Planets	4
Biology and Biochemistry	8	Stem Cells	4
Genetics and Heredity	8	Age, Chronological	3
Archaeology and Anthropology	7	Brain	3
Reproduction (Biological)	7	Cloning	3
DNA (Deoxyribonucleic Acid)	6	Earth	3
Animals	5	History	3
Diseases and Conditions	5	Mental Health and Disorders	3
Ethics	5	Religion and Churches	3
Finances	5	Universe	3
Women	5	Vaccination and Immunization	3

Table 6.9: Most frequent metadata tags in the GREAT writing samples

should be combined for assessing text quality.

6.5.2 Impact of article topic

The topic of an article invariably influences reader interest. Even at the level of news media, there is a topic bias and some science topics are more frequently reported in the news compared to others. Weitkamp (2003) notes that articles related to health and medicine are the most frequent in the science columns of British newspapers [169]. Further, among this already skewed distribution of topics, some topics grab a reader's attention to a greater extent compared to others. This section explores the degree to which topic features can accurately predict the quality categories in our corpus.

When we discussed the corpus in Chapter 3, we also listed the topic tags that are most frequently associated with articles in our GREAT category obtained from the "Best American Science Writing". We reproduce that list here in Table 6.9.

Certain topics definitely have a much higher likelihood of being interesting to users

compared to others. Medicine, Space, Physics and Biology topics are common in the GREAT writing. In comparison, History and Computer related articles are not that frequently chosen in the "Best American Science Writing".

To understand the influence of topic, we examine the use of topic tags as features for quality prediction. We take all the topic tags appearing in our corpus and remove those appearing in less than 50 articles. From the remaining set, we also ignore the 14 research tags which we used to identify the "science" related articles for corpus creation (see Table 3.2). We remove these tags since they were the topics used to create our corpus. There were 626 remaining tags which we use as features. The presence of each tag in an article is a binary valued feature.

Since topic tags are not always available on all datasets, we also approximate topic using word features. In content-based recommendation systems [108, 119], words are the standard features used to identify interesting articles for a particular user. The words from articles that the user previously read are used as indicators of user interest and other articles containing similar words are provided as suggestions. We similarly include word features to separate out the writing of GREAT authors from TYPICAL writing in the newspaper. We compute these features as follows. We identify the most frequent *1000* words in our corpus. This list is obtained after removing the 50 most frequent words and also those that appear less than 25 times in the corpus. Each word is a feature and its count in the test article is the feature value. A random sample of words from the feature list is given below.

> matter, series, wear, nation, account, chip, investor, surgery, high, receive, remember, support, worry, enough, office, prevent, biggest, customer, fear, symptom

For both tag and word features, we do not use a SVM classifier because of the large feature set size. Rather we adopt the Naive Bayes method which is standardly used along with word features in recommendation systems. The results are shown in Table 6.10.

The content features give 72% accuracy for the 'any-topic' task which is close but lower than the full set of writing features from the previous section (Table 6.6). The tag and approximate word features give similar performance for this setup. For the 'same-topic' task, the accuracies are low as expected. These examples have been normalized

Features	Any-topic	Same-topic
Tags	71.6	51.5
Words	72.8	66.5

Table 6.10: Accuracy of topic-related features

for topic so as to explore writing differences not based on topic. The tag features have close to baseline performance, only 52%. The word features since they are fine grained are stronger and provide 66.5%. Note that the full set of writing features in the previous section gave an accuracy of 78% for this task.

These results indicate that topics also give indications of interesting articles as expected and they can be well-approximated using word features. However, for the topic normalized case which is useful for search and information retrieval systems, the power of these features is much lower while the writing features give superior performance. In fact, the accuracies obtained by topic features on the 'any-topic' task could also be an optimistic estimate on this particular corpus. The reason is that we created the VERY GOOD and TYPICAL categories based on author identity. A science writer is likely to write mostly on the same topic, for example, one journalist may cover health news and another space-related research. Therefore the 'any-topic' setup might be an easier one to distinguish based on topic.

6.6 Future work

Our work is the first to perform text quality prediction for the science journalism genre. Our system can be strengthened in many ways.

Firstly, we have introduced features based only on six facets which we expected are related to quality. However, we noticed that many other aspects which are (potentially) related to writing quality such as humour, metaphor and suspense are also frequently present in science news articles. An example comparison provided in one of the science articles is below.

> Dr. Fotini Markopoulou Kalamara of the Perimeter Institute described time as, if

> not an illusion, an approximation, "a bit like the way you can see the river flow in a smooth way even though the individual water molecules follow much more complicated patterns."

Features related to such aspects can be expected to improve the accuracy of identifying the interesting articles. In future work, we plan to use ideas from computational methods to identify metaphors [46, 146], figurative language [9] and humour [106] to develop features for text quality prediction.

On the other hand, the features that we have currently implemented are quite reliable as shown by the annotation study. These features could be useful for tasks outside text quality. For example, Leong et. al [82] perform expansion of queries posed to an image retrieval system by also incorporating visual information. They suppose that visual words are likely to be better candidates for expanding such queries compared to any related word. For each candidate word they compute a picturability score using the co-occurrence of the word with image tags from the Flickr[33] database, another large corpus of images and tags. Some other recent studies have also utilized the large corpora of images and captions for obtaining visual words. Dodge et. al [36] create a dataset of visual nouns and adjectives through bootstrapping and label propagation procedures. They also use the Flickr dataset. Starting with a seed set of visual words, they discover other visual words through co-occurrence links. Their data contains close to 20,000 visual nouns and adjectives but their lexicons were not evaluated for accuracy of the markings.

Other types of evaluations are also necessary to understand the usefulness of the features we have proposed in this work. We have assumed that the six facets we studied are quite relevant for science journalism. But some of them could also be relevant for text quality prediction in other genres, for example, unusual phrasing, visual language and narrative structure. We did not have another corpus with text quality ratings to explore the genre-specific nature of our features but such an analysis would be quite interesting to perform in future. In addition, for comparison we used readability features and features for well-written nature but we trained them on our corpus categories. It would also be revealing to understand the extent to which actual readability or well-

[33] www.flickr.com/

written nature ratings influence quality prediction. To demonstrate this finding, we would need to train the readability features on a suitable corpus with difficulty ratings, for example, educational grade level marked text, and similarly for features related to well-written text. Examining the predictions of these models on our corpus will give a better view of the different aspects of quality and how their predictions differ.

6.7 Conclusions

In this chapter, we presented experiments on our new corpus of science journalism specifically aiming to develop genre-specific features related to reader interest. Compared to the effect of features from other chapters, the general-specific nature and intentional structure, the interest-related features are much more strongly predictive of quality categories.

These experiments help to motivate that by exploiting the distinctive properties of a genre, we can develop features which obtain good accuracies and also give an understanding of aspects that are indicative of article quality. To facilitate such analyses, we used a small set of intuitive features and also validated them for aspects that they represent. In contrast, previous work on predicting interesting articles, McIntyre and Lapata (2009) [104], only simple token counts were used as features. We also demonstrated more directly that interest-related features complement those proposed for other quality aspects. In McIntyre and Lapata's work, this result was only indirectly obtained—stories that were generated by their system using both entity grid and interest scores simultaneously were most preferred during evaluation by people.

Chapter 7

A model of verbosity

> *The writer is selective, avoiding trivia, and choosing details that keep the readers reading.*
>
> *The amount of detail is just right–not skimpy, not overwhelming*
>
> - Ideas and Development trait (Section 2.1)

An article is verbose when it contains unnecessary details which make the article longer than it needs to be. Such writing is unpleasing to readers. Articles that contain too much general content are also of lower quality. Overly general information conveys less meaning. So even after reading a long, very general article, a reader does not obtain much useful detail. In contrast to verbose and overly general articles, concise articles contain the right amount of detail and details which are most necessary for a reader to know. The two definitions at the beginning of this section are taken from the *Ideas and Development* category of the Six Traits model and emphasize exactly this quality aspect. In this chapter, we develop a computational method to predict verbosity and test its usefulness for making assessments of text quality.

The simple definition of verbosity is "too many words than necessary". But in a more specific sense, verbosity arises when any or both of the following factors are present (based on definitions from Williams (1990) [170]):

Redundant information: For example, in the phrase "during that period of time", the use of both 'period' and 'time' creates redundancy. This phrase could be simply written as

"during that time" or "during that period". Similarly, "terrible tragedy" can be shortened as "tragedy". This type of verbosity arises from excessive use of modifiers, complicated words and cliché phrases (eg. "each and every"). Such texts can be rewritten into concise ones such as the above examples without loss of information.

Irrelevant details: are those which the reader can infer easily and so they need not be explicit in the text. Consider the following verbose passage and its simpler concise version taken from Williams (1990)[170].

> A. Baseball, one of our oldest and most popular outdoor summer sports in terms of total attendance at ball parks and viewing on television, has the kind of rhythm of play on the field that alternates between the players' passively waiting with no action taking place between the pitches to the batter and exploding into action when the batter hits a pitched ball to one of the players and he fields it.
>
> B. Baseball has a rhythm that alternates between waiting and explosive action.

Text A is filled with unnecessary detail, for example, in the clauses, "play on the field" and "when the batter hits a pitched ball to one of the players and he fields it". In addition, depending on the reader, several other pieces of information such as the note about oldest and popular outdoor sport will also become unncessary. The rewritten text is an example concise version of the same content and brings out the main substance of the sentence. In the case of redundant information category, we can create a concise version without losing any detail whereas here, rewriting the text concisely involves conveying less information. But the content that is ignored is not important for the author's point.

The problem of verbosity in writing is often discussed in writing advice books but there have been no previous attempts to automatically predict verbosity. The work reported in this chapter is one of the first studies to propose a measurable indicator of verbosity.

This line of research is related to text specificity which we discussed in Chapter 5. In that study, we distinguished between two categories, more detail (specific) and less detail (general). Using the confidence values from the classifier, we developed a measure to indicate the level of specificity for a text. We showed in our experiments that for the task of summarization, automatic summaries with greater general content had better scores

during evaluation by human judges. Summaries with lesser general content scored lower. Similarly, when comparing articles from our science journalism corpus, the better articles according to our gold standard quality categories were those with greater general content compared to the typical articles. However, in that work, we did not examine whether the level of specificity is appropriate for individual texts. In other words, we did not examine whether the provided details are the right amount and most needed given the article's length. Our idea of verbosity prediction is designed to provide a way to check for the fit of details presented with the article length.

Our approach involves two factors—content type and article length. Specifically, we assume that certain content types are appropriate to be included in a text for a given length and some other content types are excessive and unnecessary detail. We utilize a collection of concise articles and learn a relationship between surface properties of the text (focusing on those which can indicate content type) and the length of the article in words. These properties include level of description approximated by phrase lengths and syntactic form, indications of semantic content by identifying which discourse relations are present, tracking amount of discussion on subtopics using continuity features, and external information about content that is considered important by people. This model captures which type of content and writing is ideal for long and short articles.

During testing, we analyze whether our model identifies the content type in the article as appropriate given the article length. Deviations from concise style will be reflected by a mismatch between the content type and length. We hypothesize that such mismatched articles will have lower quality compared to those where the content type and length have a good relationship when examined under the model of concise writing. Sections 7.1 to 7.5 describe our approach and how we implemented the model using a corpus of news summaries and news articles.

Similar to text specificity, we expected that this metric will be relevant for text quality prediction for automatic summaries. Since a target word limit is given for summary creation, a summarization system needs to decide how much detail to provide so that the summary will be perceived as contentful while at the same time not involving unncessary details. We show that our model is predictive of both content and linguistic quality ratings

for automatic summaries (Section 7.7). We also used features from our verbosity model for predicting the quality of science journalism articles. However, we did not gain any performance improvement above baseline for this task (Section 7.8).

7.1 Content type and verbosity

In this section, we explain the idea behind our approach: that we can predict whether a text is verbose or not based on the types of detail included in the text.

Verbose articles use more words than necessary. But the length of an article alone does not indicate verbosity. A long article article can be gracefully and concisely written. At the same time, a short paragraph such as snippet (A) in the previous section can include a lot of irrelevant details and be verbose. A short paragraph can also be overly general and full of meaningless cliché and modifiers.

We discussed two factors that are indicative of verbosity—redundant information and irrelevant details. Our model for verbosity is based on the second factor. In particular, we assume that:

> *For different length articles, there is an appropriate level and certain types of detail that can be included. For short articles, some types of content are appropriate. For long articles, some other types of content are suitable. Length alone does not differentiate verbose from concise writing. Rather the type of content and its suitability for the article's length is the determining factor.*

For example, consider the summaries written for the same input at two different lengths (50 and 100 words) and by the same person (Table 7.1). These examples are taken from the Document Understanding Conference dataset (year 2001) where expert assessors wrote summaries of different lengths which were then used as a gold standard for evaluating machine generated summaries. These assessors are retired information analysts and so these summaries can be considered as concise and appropriately written for the different lengths.

At an abstract level, the information conveyed by both summaries is the same but with increasing granularity and detail. For example, consider two of the facts that can

50 word summary:

The De Beers cartel has kept the diamond market stable by matching supply to demand. African nations have recently demanded better terms from the cartel. After the Soviet breakup, De Beers contracted for diamonds with the Yukutian Republic. The US remains the largest diamond market, followed by Japan.

100 word summary:

The De Beers cartel, controlled by the Oppenheimer family, controls 80% of the uncut diamond market through its Central Selling Organization. The cartel has kept the diamond market stable by maintaining a buffer pool of diamonds for matching supply to demand. De Beers opened a new mine in 1992 and extended the life of two others through underground mining. Innovations have included automated processing and bussing workers in daily from their homes. African nations have recently demanded better terms. After the Soviet breakup, De Beers contracted for diamonds with the Yukutian Republic. The US remains the largest diamond market, followed by Japan.

Table 7.1: 50 and 100 word summaries written by one person for a multidocument input

be inferred by a reader from both these summaries: a) De Beers is a diamond cartel, and b) De Beers has kept the diamond market stable by matching supply to demand. But these facts are conveyed with quite different granularity in the two summaries. In the 100 word summary we are also told that De Beers is led by the Oppenheimer family and that it also controls 80% of the diamond market. The introduction of De Beers in the 50 word summary does not contain these details. Fact (b) is directly stated in the 50 word summary. However, the 100 word summary provides more details—De Beers has created the stability by maintaining a buffer pool of diamonds. The new mine that it opened and innovations in processing has helped De Beers to do so. These causal details are not present in the shorter summary.

Our idea for developing a verbosity prediction method is based on such differences in content for articles of different lengths. Certain types of information should be omitted when writing for a shorter versus longer length. If the 50 word summary attempts to provide the kinds of details present in a 100 word summary, it would be too detailed and verbose. If a 100 word summary is written in the 50-word style, it will be overly general. For example, it is almost certain that a 1000 word article about De Beers would not have the same type of introduction as the 50 word summary. In our model, we aim to automatically learn this relationship between content type and article length using surface linguistic properties of the text. For example, based on the example summaries above, some of the surface properties related to content type are phrase lengths, sentence specificity, and discourse relations. We discuss the details of this approach in the next section.

We believe that this definition of verbosity can be directly useful for many NLP systems that need to control the length of text. Generation and summarization are good examples. Currently, summarization systems first generate a ranked list of sentences according to their importance in the article. Then starting with the sentence with highest value, the system adds the sentence to its summary before moving on to the next sentence in its list. The selection process stops when the summary reaches its desired length. This strategy places no attention to the fact, that a 50 word summary is written so differently compared to a 100 or a 400 word summary. We expect that our findings and verbosity

prediction method can help improve sentence selection and compression techniques for summarization. For example, compression can be modeled such that different types of deletions are made depending on the target summary length.

In our work, we have focused only on the aspect of verbosity dealing with the presence of irrelevant details. The identification of the other aspect—redundant information—is also important for predicting verbosity. In the automatic summarization field, there is a lot of interest in developing ways to identify and remove redundant content from summaries [15, 143]. In future, we plan to explore how such techniques can be combined with our approach for verbosity prediction. We hypothesize that other specialized techniques will also be necessary to identify redundancy arising from use of cliché phrases and meaningless and redundant modifiers.

7.2 Model summary

In this section, we explain how we learn the relationship between content type and article length and how we use this model during test time to predict verbosity.

We first define some assumptions that we make:

1. **Information content and length:** Let us assume that under optimal (i.e. concise) writing, when more information needs to be conveyed, a writer would create a longer text. For example, a 100 word text conveys less information compared to a 500 word text.

2. **Variation in content types:** Content type varies accordingly depending on the length. In long concise articles, the writing is designed to convey more information compared to a short article. The types of content included is therefore different from that in short concise articles.

3. **Differences from optimal writing:** The two extremes from the optimal writing situation are—a) conveying irrelevant and excessive detail and b) conveying vague and less meaningful information.

We wish to model the dependence between content type and length using the following training approach:

- Let $D = (d_1, d_2, ...d_n)$ be a collection of concisely-written articles and let $l(d_i)$ denote the length of article d_i. The learning task is to obtain a function based on the content type properties of d_i which helps to predict $l(d_i)$. More specifically, we are given a snippet from d_i, called w_{d_i}, of a constant length k where $k < \arg\min_{d_j} l(d_j)$. The mapping f is learned based on the constant length snippet from any article and the aim is to predict its original length.

$$f(w_{d_i}) \rightarrow \hat{l}(d_i)$$

- An article would contain different information and involve different writing styles in different parts of the article. So rather than article length, we will model the length of topic segments, where a topic segment is a coherent discourse segment from an article. Therefore in the learning task, we replace D with the set T of topic segments $t_1 ... t_T$ from our corpus.

The prediction idea is as follows:

Let us consider a new topic segment t_x during test time. Let the length of the segment be l. We obtain a snippet w_{t_x} of size k from t_x. Now assume that our model predicts $f(w_{t_x}) = \hat{l}$.

- Case 1: $\hat{l} \simeq l$, the content type in t_x matches the content types generally present in articles of length l.

- Case 2: $\hat{l} \gg l$, the type of content included in t_x is really suitable for longer and detailed topic segments under concise writing scenario. So t_x may be conveying too much detail given its length.

- Case 3: $\hat{l} \ll l$, the content in t_x is of the type that an excellent writer would include in a much smaller and less detail-oriented text. So t_x could be overly general and lacking appropriate details.

We propose that the disconnect in case 2 is the closest to verbosity with irrelevant details. Case 3 indicates overly general articles.

7.3 Features for length prediction

We propose the following features for characterizing the content type of articles. These features are computed over the constant length snippet obtained from the articles. There are a total of 87 features which we added based on different motivations which we describe below. Some of the features require syntax information. We used the Stanford Parser [75] to obtain the constituency parse trees for the sentences in the snippet.

Length of units (10 features).

This set of features captures basic word and sentence length and redundancy properties of the snippet.

It includes number of sentences, average sentence length in words, average word length in characters, and type to token ratio. We also include the counts of noun phrases, verb phrases and prepositional phrases and the average length in words of these three phrase types.

Syntactic realization (30 features).

These features are based on grammatical productions obtained from constituency parse trees.

We compute the most frequent productions in a set of news articles from the AQUAINT corpus [54] (47472 sentences total). From this set, we record the top 15 productions that involve the description of entities, i.e the LHS (left-hand side) of the production is a noun phrase. The count of each of these productions is added as a feature. Table 7.2 shows these entity realization features.

Similarly we find the most frequent 15 productions whose LHS is not a noun phrase. These productions are listed in Table 7.3. The count of each of these production is also a feature.

We expect that these syntax features will capture how entities and other phrases in the snippet are realized and what kind of information is attached to them.

NP→NN	NP→NNS	NP→NNP
NP→CD	NP→NNP NNP	NP→PRP
NP→JJ NNS	NP→DT NNS	NP→DT NN
NP→DT JJ NN	NP→JJ NN	NP→NP PP
NP→NP CC NP	NP→NP SBAR	NP→NP VP

Table 7.2: Frequent productions related to entity descriptions

PP→IN NP	ROOT→S	PP→TO NP
S→NP VP	S→VP	S→NP VP .
SBAR→IN S	SBAR→WHNP S	SBAR→S
VP→TO VP	VP→VB NP	VP→MD VP
VP→VBN PP	VP→VBZ VP	ADVP→RB

Table 7.3: Frequent productions related to non-entity type phrases

Discourse relations (5 features).

These features are based on the hypothesis that different discourse relations would vary in their appropriateness for articles of different lengths. In a different study, Louis, Joshi and Nenkova (2010) [91], we examined which discourse relations are indicative of content that is selected by people for very short summaries. We found that there are some significant discourse indicators for people's preferences . For example, explicit expansion and contingency relations are significantly less preferred to be included in short summaries compared to other discourse relations. We expected that similar differences may exist for the distribution of discourse relations in articles of different lengths.

We run the *addDiscourse* tool[34] developed by Pitler and Nenkova (2009) [125] to identify the explicit discourse relations in our snippets. Explicit relations are those which are signalled through the presence of a discourse connective such as 'because', 'but' or 'after'. The tool is trained on the Penn Discourse Treebank [128] annotations and marks every connective as indicating one of four discourse relations—Comparison, Contingency, Expansion and Temporal. The features are the counts of each of the four types of relations as well was the total count of all relations.

[34] http://www.cis.upenn.edu/~epitler/discourse.html

Continuity (6 features).

These features capture the degree to which adjacent sentences in the snippet are related. High similarity between adjacent sentences could indicate continued discussion of a subtopic. On the other hand, low similarity can signal change of topic. In short articles, not much space is available for detailed discussion of a subtopic while such details can be provided in longer articles. So we expect continuity to also indicate content and information differences in the articles.

For this purpose, we include the number of pronouns and determiners as two features. We also include average word overlap value between adjacent sentences. For computing the overlap measure, we represent every sentence as a vector where each dimension represents a word. The count of the word in the sentence is the value for that dimension. Cosine similarity is computed between the vectors of adjacent sentences and the average value of the similarity across all pairs of adjacent sentences is a feature.

We also run the Stanford Coreference tool [136] to identity pronoun and entity coreference links within the snippet. We add the total coreference links as a feature, and the total number of intra-sentence and inter-sentence links as additional features.

Amount of detail (7 features).

We quantify the amount of detail using features from our general-specific classifier. We add two features, the percentage of specific sentences and the average specificity of words (see Chapter 5 Section 5.5.2).

We also add the number of descriptive words such as adjectives and adverbs (two features). To indicate specific details, we also include the total number of named entities (NEs), average length of NEs in words and the number of sentences that do not have any NEs. The named entities were identified using the Stanford Named Entity Recognition tool [48].

Compression likelihood (29 features).

These features use an external source of information about content importance.

Specifically, we use data that is commonly employed to develop statistical models for sentence compression [52, 76, 103]. In such studies, the training data consists of pairs of sentences: one sentence comes from an abstract (summary) written for an article

and the other sentence comes from the actual article and has close content similarity with the abstract sentence. (These summaries are written by people and are not system generated.) The short sentence is assumed to be a compression of the source article sentence created by the summary writer. Further, since the sentence was shortened for inclusion in a summary, it is assumed that the summary sentence has the most important content and extraneous details are removed by the writer during compression. The goal of the sentence compression task is to automatically perform the transformation from the source to shorter summary sentence.

The Ziff Davis corpus [63] has been commonly used for creating datasets for compression experiments. It contains articles about technology products and every article includes a summary. We use the mappings of source and abstract sentences created by Galley and McKeown (2007) [52]. They allowed a mapping when any number of words from the source can be deleted and upto 7 substitutions operations can transform the source to the shorter abstract sentence. This data also contains alignment between the constituency parse nodes of the source and abstract sentence pair. The alignment indicates which nodes from the source were preserved in the abstract sentence.

Using this data, we identify for every production in the source sentence whether it undergoes deletion in the abstract sentence. A deletion is defined as follows: for a production LHS $\rightarrow$ RHS, when either the LHS node or any of the nodes in the RHS do not appear in the abstract sentence, we consider that a deletion has been made within that production. Only productions which involve non-terminals in the RHS are used for this analysis. Lexical items could be corpus specific and not likely to generalize to other datasets. So we ignore them.

The proportion of times a production undergoes deletion is called the *deletion probability*. We also incorporate frequency of the production with the deletion probability to obtain a good representative set of productions which are frequently deleted and also occur commonly. This *deletion score* is computed as:

$$\text{deletion probability} * \log(\text{frequency of production in source articles})$$

The 25 productions with highest deletion scores are shown in Table 7.4.

S→S , NP VP .	NP→NP , NP	VP→VP CC VP	PRN→LRB NP RRB
S→PP , NP VP .	NP→NP PP PP	VP→VB NP PP	PP→VBG PP
S→S , CC S .	NP→NP , SBAR ,	VP→VP , CC VP	PP→IN S
S→S : S .	NP→NP , NP , CC NP	VP→VBD	WHNP→WDT
S→ADVP , NP VP .	NP→NP : NP		
S→SBAR , NP VP .	NP→DT		
S→NP ADVP VP	NP→NP PRN		
	NP→NP , SBAR		
	NP→NP , NP ,		
	NP→NP NP		

Table 7.4: Most deleted 25 productions from the Ziff Davis Corpus

We find that parentheticals appear in the list as would be expected and also productions involving conjunctions and prepositional phrases. We expect that such productions will indicate how much detail is present which potentially is less important and likely to be deleted while creating a summary of the article.

The features for a snippet are computed as follows. We obtain the set of all productions in the sentences from the snippet. We add the sum, average and product of deletion probabilities for the productions as features. The product feature gives the likelihood of the sentence being deleted. We also add the perplexity value based on this likelihood, $P^{-1/n}$ where P is the likelihood and n is the number of productions from the snippet for which we have deletion information in our data[35].

We also add the frequency of each production in the most deleted set above as a feature.

For training a model, we need texts which we can assume are written in a concise manner. We use two sources of data—summaries written by people and high quality news articles. These datasets and the training approach is detailed in the next two sections.

[35]Some productions may not have appeared in the Ziff Davis Corpus.

7.4 A classification model on expert summaries

In this experiment, we use a collection of summaries written by expert people. The summaries were created for four lengths. So we build a classification model on this data to predict given a snippet what is the length of the summary from which the snippet was taken. This task is simple and limited, and only differentiates four lengths. However, it is a useful first approach for testing our assumptions and features.

7.4.1 Data

Our summaries come from the Document Understanding Conference (DUC[36]) evaluation workshops conducted in 2001 and 2002. In these first two years of the workshop, the task for automatic systems was to create summaries of different lengths. The input given to systems was a set of 10 to 15 documents on the same topic. The systems had to create 50, 100, 200 and 400 word summaries for each of the inputs. To evaluate the automatic systems, assessors at NIST (which conducts the DUC evaluations) also wrote summaries at these four lengths for all the inputs. These assessors are retired information analysts who had worked for the government and are experts in writing summaries. Therefore we can assume that their summaries are of high quality and concise and informative nature. Our example summaries used in Section 7.1 also come from the same data.

Further, apart from the fact that summaries at different lengths are available, there are two additional advantages of this dataset. One advantage is that the four different length summaries for an input are produced by the same person. (Different inputs however may be summarized by different assessors.) Therefore differences in length are not confounded by differences in writing style of different people. An additional advantage is that overall the summaries come from a few inputs (30 in total), so there are only a small set of topics corresponding to the texts in the dataset. Also, summaries on each topic are available for all the four lengths. Such a setting allows us to examine changes in content type depending on the length without much concern about how the content or writing varies depending on the topic.

The 2001 dataset had 30 inputs and for each we have 3 summaries of each length.

[36] http://duc.nist.gov

Therefore there are total of 90 summaries for each of the four lengths. In 2002, there are 59 inputs and 2 summaries each for the four lengths. This gives us 116 summaries for each length in the 2002 data. All of the summaries are abstracts, people wrote the summary in their own words, with the exception of one. In 2002, abstracts were only created for 50, 100 and 200 lengths. However, extracts created by people are available for 400 words. In extracts, there is less flexibility given to the person creating the summary. He is only allowed to choose complete sentences from the input for including in the summary and no edits are done to individual sentences. However, the sentences can be ordered in the summary and people tend to create quite coherent summaries under the extract condition as well. Since it would be nice to have data corresponding to another length as well, we include these extracts as the 400 word class of summaries in the 2002 data.

7.4.2 Snippet selection

We chose 50 words as the snippet length for our experiment since the length of the shortest summaries is 50. Since the content and writing would vary in different portions of an article, we experiment with multiple ways to select a snippet: the first 50 words of the summary (START), the last 50 words (END) and 50 words starting at a randomly chosen sentence (RANDOM). However, we do not truncate any sentence in the middle to meet the constraint for 50 words. We allow a leeway of 20 words so that snippets can range from 30 to 70 words. When a snippet could not be created within this word limit (eg. the summary has one sentence which is longer than 70 words), we ignore the example.

7.4.3 Classification results

We used the 2001 data for training a classifier and test the classifier on the 2002 data. The task is a 4-way classification, between whether the snippet came from a 50, 100, 200 or a 400 word summary. We trained a SVM classifier with a radial basis kernel. The regularization and kernel parameters were tuned using 10-fold cross validation on the training set. We then ran the model on the 2002 data and the accuracies of classification are show in Table 7.5. Since there are four equal classes, the random baseline performance is 25%.

snippet position	accuracy
START	38.4
RANDOM	34.4
END	39.3

Table 7.5: Accuracies for predicting length on DUC summaries

The START and END position snippets gave the best accuracies, 38% and 39% which are 13-14% above the baseline. These results indicate that the model is gives significantly good performance however there is much scope for improvement. We analyse these results further by looking at the confusion matrices for the START and END snippet positions (Tables 7.6 and 7.7).

We find that 50 and 400 word lengths, the extreme ones in this dataset, have been the easiest to predict. About 50% or more of the examples in these lengths are classified correctly. Most of the confusions occur with the 100 and 200 word summaries. At least in the START snippet selection case, we find that the confusions for 100 and 200 word summaries are high with other classes of closer length than those farther away. But a unusual result is that for 50 and 400 word summaries, there are high confusions with each other. It could indicate that these summaries have special characteristics that are indicative of their class, but mistakes are spread out over the other classes. We also expect that an approach that takes into account the fact that, these classes have an order, 50 is less than 100 which in turn is less than 200, would provide better accuracies.

The overall accuracy is slightly better when snippets from the END of the summary are chosen compared to those from the START. However, in the START selection case, better prediction of different classes of summaries are obtained (including 200 word summaries) whereas the accuracy in the END case comes mainly from correct prediction of 50 and 400 word summaries. So we use the START selection for further experiments.

		PREDICTED LENGTH				
		50-word	**100-word**	**200-word**	**400-word**	**Total summaries**
TRUE LENGTH	**50-word**	66 (56.9)	19 (16.3)	15 (12.9)	16 (13.8)	116
	100-word	39 (34.2)	27 (23.7)	31 (27.2)	17 (14.9)	114
	200-word	34 (29.8)	22 (19.3)	24 (21.1)	34 (29.8)	114
	400-word	18 (16.5)	9 (8.3)	21 (19.3)	61(55.9)	109

Table 7.6: Confusion matrix for length prediction with END snippet selection on DUC summaries. The counts are also normalized by the total number of summaries of each length (in the last column) and shown within parentheses. (The number of summaries varies for different lengths because any summary where a suitable 50 word snippet could not be obtained was ignored. See Section 7.4.2.)

		PREDICTED LENGTH				
		50-word	**100-word**	**200-word**	**400-word**	**Total summaries**
TRUE LENGTH	**50-word**	65 (56.0)	12 (10.3)	18 (15.5)	21 (18.1)	116
	100-word	43 (37.4)	17 (14.8)	26 (22.6)	29 (25.2)	115
	200-word	24 (21.2)	14 (12.4)	43 (38.1)	32 (28.3)	113
	400-word	27 (24.1)	14 (12.5)	21 (18.8)	50 (44.6)	112

Table 7.7: Confusion matrix for length prediction with START snippet selection on DUC summaries.

7.5 A regression approach based on New York Times editorials

Based on the success with the classification approach, we move to a model where we accomodate the prediction of a wider range of lengths compared to just the four classes we had before. For these we use news articles from the New York Times which are of high quality overall. This model uses a linear regression method on the actual lengths of the articles. These experiments are detailed in this section.

7.5.1 Data

We noticed that general political news has less continuity between its sentences and paragraphs. They commonly follow the 'inverted pyramid' structure of news reporting [126] where a summary of the event is provided in the beginning of the article and then the information is arranged in order of decreasing importance. This structure is not ideal for our model which uses a lot of discourse and continuity features. So we choose to use articles from the opinion section of the newspaper. To comply with our assumption that the articles are overall of optimal quality, we further use only the editorial articles published in this section. Especially in the New York Times, the editorial board is comprised of outstanding journalists including Pulitzer prize winners. So we can expect that these articles are of very good quality overall.

We collect the editorial articles in the opinion section from 2000 to 2007 years of the New York Times. We use the metadata in the NYT corpus [142] for identifying the target articles and obtaining their full text. This selection provides us with 10,734 articles. Some articles are very short, 30 to 100 words and they are often letters. The maximum length of articles is about 2500 words.

7.5.2 Training approach

We divide each article into topic segments since an entire article is unlikely to have uniform content and writing. For this purpose, we use the unsupervised topic segmentation model developed by Eisenstein and Barzilay (2008) [39]. We use the following heuristic to decide on the number of topic segments for each article. If the article has less than 50 sentences, we create segments such that approximately 10 sentences have the opportunity

to be in a segment, ie, we assign the number of segments as number of sentences divided by 10. When the article is longer than that length, we create 5 segments. We also remove articles that have less than 10 sentences. This step gives us 18,167 topic segments.

These topic segments also have varied lengths, from 14 to 773 words. In order that the model learns the content type properties related to different lengths and because the distribution of lengths for topic segments may not be the same on different corpora, we use a stratified sampling method to select training and test examples. Starting from 90 words and upto a maximum length of 500 words, we divide the range into bins for every 30 words. There are a total of 21 such bins. From each bin we select 100 texts for training and around 35 for testing. There are 2,100 topic segments in the training set and 681 for testing.

We choose a length of 100 words for snippet creation on this data. We compute the features on the training set and train a linear regression model on the data. We use the *lm* function within R [132] to perform the regression. The features which turned out significant in the model are shown in Table 7.8. The significance value shown is associated with a t-test to determine if the feature can be ignored from the model. These features are most important for the fit of the model. We report the coefficients for the significant features under column 'Beta'. We show features from different levels of significance, p-value less than 0.001 to p-value less than 0.1. The R-squared value of the model is 0.219.

7.5.3 Accuracy of predictions

On the test data, the lengths predicted by the model have a correlation of 0.44 with the true length of the topic segment. The correlation is highly significant with a p-value less 2.2e-16. The plot of these two lengths is shown in Figure 7.1. We find that the correlation is not very strong but there is a general trend in the predictions. The Spearman correlation between the lengths is 0.43 and the Kendall Tau is 0.29, both also highly significant.

This result indicates that our approach and features are useful for predicting length at the finer level as well.

We also ran the regression model on the test summary data from DUC (year 2002). Since the regression model uses 100 words, we use 100 word snippets on the DUC test set

Feature	Beta	p-value	Feature	Beta	p-value
Positive coefficients			**Negative coefficients**		
total noun phrases	6.052e+00	***	NP → NNP	-8.630e+00	***
avg. word length	3.201e+01	***	no. of sentences	-2.498e+01	**
avg. sent. length	3.430e+00	**	no. of relations	-1.128e+01	**
avg. NP length	6.557e+00	*	avg. VP length	-2.982e+00	**
no. of adverbs	4.244e+00	**	type token ratio	-1.784e+02	*
% specific sentences	4.773e+01	**	NP → NP , SBAR	-1.567e+01	*
comparison relations	9.296e+00	.	NP → NP , NP	-9.582e+00	*
determiners	2.955e+00	*	NP → DT NN	-3.423e+00	.
NP → NP PP	4.305e+00	*	VP → VBD	-1.189e+01	.
NP → NP NP	1.174e+01	*	S → S : S .	-1.951e+01	.
PP → IN S	7.268e+00	.	ADVP → RB	-4.198e+00	.
WHNP → WDT	1.196e+01	**			

Table 7.8: Significant regression coefficients in the length prediction model on NYT editorials. '***' indicates p-value < 0.001, '**' is p-value < 0.01, '*' is < 0.05 and '.' is < 0.1

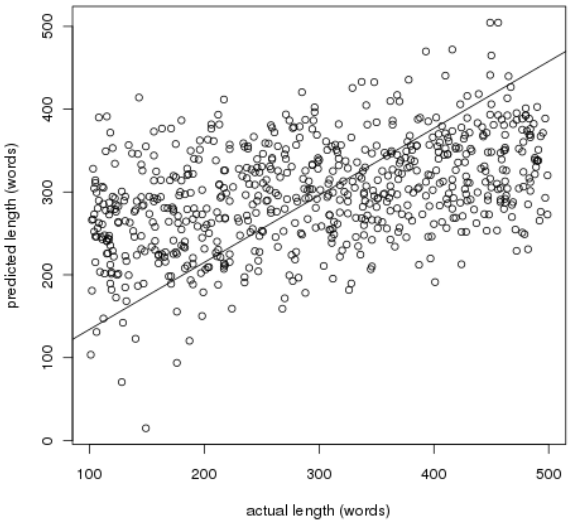

Figure 7.1: Plot of actual topic segment length of NYT articles and the predicted length under the model

	Predicted length			
True length	**Min**	**Max**	**Mean**	**Median**
100	-199.7	725.7	245.9	250.3
200	119.8	537.7	264.5	265.3
400	69.8	437.6	311.3	316.6

Table 7.9: Predictions from the NYT regression model on the DUC 2002 data

as well. So only 100, 200 and 400 word summaries are used. The minimum, maximum, mean and median value of the predicted lengths on these three types of summaries is shown in Table 7.9.

These results show that the predictions from the regression approach are generalizable beyond just the specific test set from opinion articles. The mean and median values of the predicted lengths are clearly higher with greater actual summary lengths. However, these predictions are not the same as the true lengths of the summaries. A two-sided t-test showed that the predictions on 100 and 200 word summaries are significantly lower than 400. The 100 word summaries have lower value predictions than on 200 word summaries, the p-value here is close to significance (0.054).

7.6 An application of the predictions to analyze literary texts

So far we developed our models by assuming that the writing in the DUC summaries and New York Times editorials is of concise and well-written nature. It would be interesting to test how this model performs for distinguishing actual examples of verbose and non-verbose writing. However, such gold standards are not available. It is also an interesting issue to consider if people can provide direct ratings for verbosity of text. It is likely that people could point to the verbose version when given a pair of texts but individual scoring is likely to be a hard task. We plan to explore annotation of verbosity in future work. Rather in this section, we provide a simple analysis of our model's predictions on literary articles which are known to come from wordy and very concise styles.

Specifically, we take writing samples of two authors, Charles Dickens and Ernest Hem-

Charles Dickens	**Ernest Hemingway**
Great Expectations	The Short Happy Life of Francis Macomber
Bleak House	The Capital of the World
A Tale of Two Cities	Snows of Kilimanjaro
A Christmas Carol	A Natural History of the Dead
David Copperfield	Big Two-Hearted River
The Old Curiosity Shop	My Old Man

Table 7.10: Novels and stories selected for the studying the verbosity model

ingway. These two authors are known for their distinctive writing styles: Dickens for his wordy and lengthy novels and Hemingway for his remarkably succinct and concise style of writing. We obtain the text of a few novels of Dickens and a set of short stories from Hemingway and check how our model's predictions vary for these texts. In addition to the writing styles of the authors, note that the novel/short story distinction is also present in the data. Novels could exhibit a more verbose style compared to short stores.

7.6.1 Data

We obtain the text of the stories from Project Gutenberg[37] and `archive.org` websites. The texts collected for the two authors are shown in Table 7.10.

In order to obtain writing samples from different part of the articles, we collect 10 *samples* of 1000 consecutive words from each novel. We create five topic segments on each sample. Then for each topic segment, using the first 100 words as the constant snippet, we try to predict the expected length of the topic segment. Then we add up these predicted lengths of the five topic segments for a sample. If the content type of each topic segment matched its true length, then the sum of the predicted lengths for the five segments should be approximately 1000. High or low values indicate deviations from concise writing as defined by the training data.

[37]http://www.gutenberg.org/

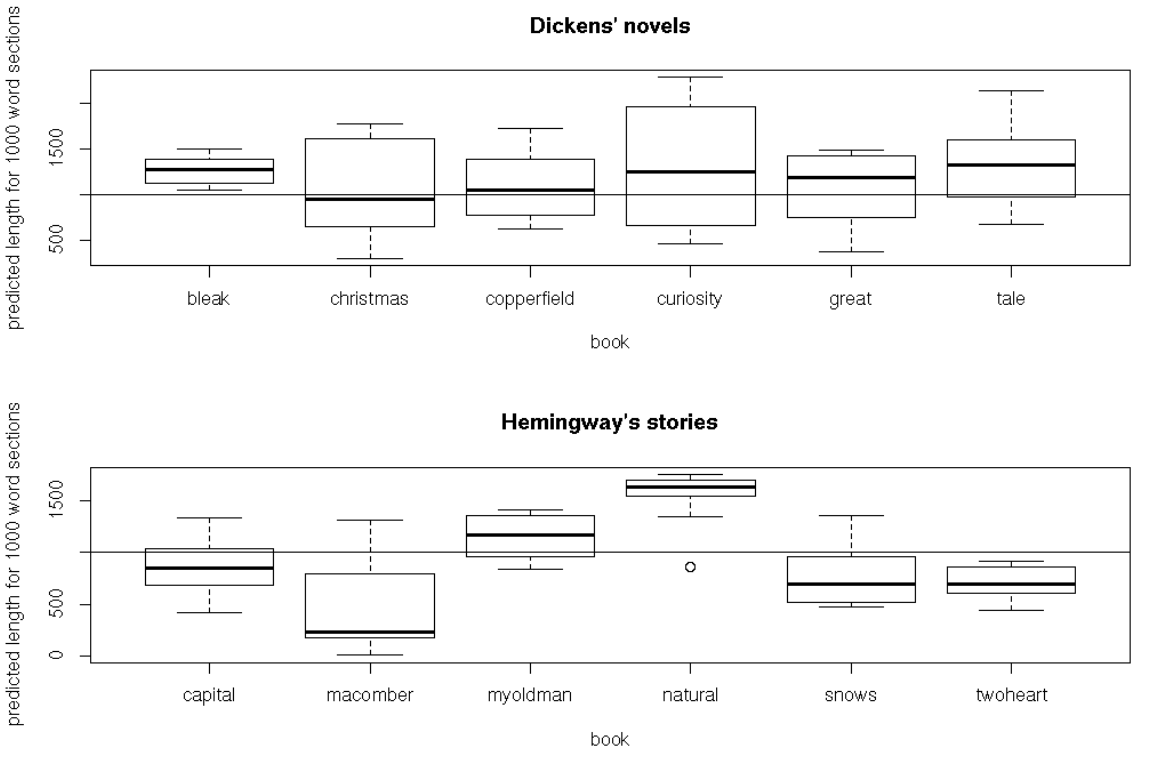

Figure 7.2: Predicted lengths for 1000 word samples from Dickens' and Hemingway's writing. The horizontal line indicates the 1000 word mark.

7.6.2 Prediction results

We use the NYT regression model to do the prediction. For each topic segment, we use the first 100 words as the snippet to examine. Sometimes, the topic segments were such that the size was much less than 100, in these cases, we ignore the topic segment and do not make any predictions for it. Since these segments are small anyway, they do not affect the actual total length of 1000 words which the predictions should approximate.

The summed up values for each of the 10 samples are expected to be close to 1000 if the concise style just like in the NYT is followed. However, since these texts are from a much different domain, even at the outset, some differences are expected. Figures 7.2 shows the distribution of predicted values for the 10 samples from each novel.

We find that the model reasonably distinguishes between the two styles of writing. Most of Dickens' novels have a median predicted length of either 1000 or more while for

While it is, perhaps, legitimate to deal with these self - designated citizens in a natural history of the dead, even though the designation may mean nothing by the time this work is published, yet it is unfair to the other dead, who were not dead in their youth of choice, who owned no magazines, many of whom had doubtless never even read a review, that one has seen in the hot weather with a half-pint of maggots working where their mouths have been. It was not always hot weather for the dead, much of the time it was the rain that washed them clean when they lay in it and made the earth soft when they were buried in it and sometimes then kept on until the earth was mud and washed them out and you had to bury them again.

Table 7.11: Example snippet ("A Natural History of the Dead") which was predicted with much greater length than actual

Hemingway's writing, they are mostly below the 1000 line mark. There are two exceptions in Hemingway's writing, "My Old Man" and "A Natural History of the Dead".

We noticed that the latter does have a rather unusual style compared to other Hemingway's stories. It is more essay like and describes philosophicals views of how people think about death. An example snippet from "A Natural History of the Dead" is shown in Table 7.11. Its length is 142 words but our model predicted that this topic segment has content which is typically used in a 419 word article (or topic segment in this case).

Table 7.12 gives example snippets (only first few words of the topic segments) from Dickens writing where the content type is predicted to be suitable for a much shorter length or for a much longer length passage.

7.7 Text quality assessment for automatic summaries

Now we turn to text quality predictions based on this model for the genres in our work. In this section we detail experiments on assessing summary quality. We perform this evaluation for the system summaries produced during the 2006 DUC evaluation workshop.

Our data consists of 20 multidocument inputs. Each input is a set of 25 documents

Predicted as much longer

Then up rose Mrs. Cratchit , Cratchit 's wife, dressed out but poorly in a twice-turned gown, but brave in ribbons, which are cheap and make a goodly show for sixpence; and she laid the cloth, assisted by Belinda Cratchit, second of her daughters , also brave in ribbons; while Master Peter Cratchit plunged a fork into the saucepan of potatoes, and getting the corners of his monstrous shirt collar (Bob's private property, conferred upon his son and heir in honor of the day) into his mouth, rejoiced to find himself so gallantly attired, and yearned to show his linen in the fashionable Parks.

Predicted as much shorter

'What's to-day!" cried Scrooge, calling downward to a boy in Sunday clothes, who perhaps had loitered in to look about him.

"EH ?" returned the boy, with all his might of wonder.

"What's to-day, my fine fellow?" said Scrooge.

"Today!" replied the boy.

"Why, CHRISTMAS DAY."

"It's Christmas day!", said Scrooge to himself.

Table 7.12: Example snippets ("A Christmas Carol") which had much deviation from actual length

on a topic. The task given to systems is to produce a summary of 250 words for each input. There are 22 automatic systems in that dataset. (We use only the set of systems for which pyramid scores are also available.) Each system produced a summary for the inputs and they were evaluated by DUC assessors for multiple dimensions of quality. We examine how the predictions from our model are related to these summary scores in the DUC data. In this experiment, we use automatic summaries only.

7.7.1 Gold-standard summary scores

There were two kinds of scores—content and linguistic quality—provided to each summary during the DUC evaluation.

For content, two different scores were assigned. One is called the 'pyramid score' [114] which is computed by comparing the semantic units of the system summary to summaries created by people. Moreover the comparison is done with summaries created by multiple people for the same input so as to understand which content is most important and so mentioned by several people in their summaries. The system summaries which have high overlap of their units with human summaries receive a higher pyramid score. The other content score is called 'content responsiveness'. For this score, assessors directly provide a rating to summaries on a scale from 1 (very poor) to 5 (very good) based only on content quality.

Linguistic quality is evaluated separately on the basis of few quality questions. There are typically 5 questions one for each aspect—grammar, non-redundancy, focus, coherence and referential clarity. For examining our verbosity model, non-redundancy, focus and coherence scores for summaries appear most relevant. For each aspect, the summary is rated by NIST assessors on a scale from 1 (very poor) to 5 (very good). The definition of these aspects as given to the assessors is given below.

> **Non-redundancy:** There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., "Bill Clinton") when a pronoun ("he") would suffice.
>
> **Focus:** The summary should have a focus; sentences should only contain information that is related to the rest of the summary.

Structure and Coherence: The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

This summary data and scores is quite useful and a good gold-standard for assessment. However, it is less ideal in some ways. Summaries produced by systems are often extractive and created by selecting important source document sentences. When these sentences are chosen from multiple source documents, they have very low cohesion and coherent arrangement in the summaries. Systems also rarely make an attempt to order the sentences. Our model and features rely on coreference and adjacent sentence overlaps and therefore some of the statistics computed on these summaries could be misleading. However, with this caveat, the summary data is the best one available at this time.

7.7.2 Verbosity scores from our model

We computed different scores based on our model. We chose a snippet size of 100 words and used the NYT regression model to predict the expected lengths of these summaries. No topic segmentation is done and the first 100 words of the summaries are taken as the snippet. From these predictions, we obtained three types of scores:

VScore1: Predicted length. This is the expected length given the type of content present in the summary. According to our assumption longer lengths are given to texts whose content type is suitable for writing longer articles.

VScore2: Verbosity degree. This score is the difference between the predicted length and the actual length of the summary. Although the summaries are all supposed to be 250 words in length (according to the task description), some of them are longer or shorter, ranging from 100 to 280 words under our tokenization method. So we compute this score as (predicted length - actual length of summary)

Vscore3: Deviation score. If the predicted length is less than the actual length the verbosity degree score is very low and negative. However, using overly general content in a long actual article could also be detrimental to quality. So we also compute a score to indicate any deviation of the predicted length from actual length. This score is given by the absolute magnitude | predicted length - actual length |.

Actual length. To understand how these verbosity scores are related to the length of the summary, we also keep track of the actual number of words in each summary.

Then for each of the 22 automatic systems, the scores of its 20 summaries (one for each input) are averaged. (We ignore empty summaries and those which are much smaller than the 100 word snippet that we require). We find the average values for both our verbosity based scores above and the gold-standard scores (pyramid, content responsiveness, focus, non-redundancy and coherence). We also compute the average value of the summary lengths for each system.

First we examined how the verbosity scores are related to the actual summary lengths. The correlations are below (Table 7.13).

Verbosity scores	**Correlation with actual length**
predicted length	-0.01
verbosity degree	-0.29
deviation score	-0.27

Table 7.13: Relationship between verbosity scores and summary length

We find that the verbosity scores are not significantly related to summary length. They seem to have an inverse relationship but the correlations are not significant even at 90% confidence level. We would expect to find a high correlation between predicted and actual lengths when the summaries are all concise as detected by our model. Here since the summaries are produced by automatic systems, they are unlikely to have such well-written characteristics. The fact that predicted length is not correlated with the actual one could be indicative of the low quality of summaries but this result gives no direct validation of that claim. The analysis between these scores and the summary quality measures which we examine in the next section is needed for this conclusion.

Tables 7.14 and 7.15 show two summaries produced for the same input by two different systems. They both have almost the same actual length but the first received a prediction close to its actual length while the other is predicted with a much higher verbosity degree score (as defined above). These examples give an idea of the distinction that our scores make on the test examples. Intuitively, the second example does appear

more verbose compared to the first one.

7.7.3 Redundancy score

In the introduction to this chapter and in Section 7.1, we noted that there are two factors which contribute to verbosity of text—1) redundant information and 2) irrelevant details. We proposed that our approach which we have implemented is based on the second aspect of whether the details are necessary for a text given its length. But redundancy is also an important component of verbosity.

So we also add a simple score to our analysis to indicate redundancy between adjacent sentences in the summary. We represent each sentence using a vector. Each dimension in the vector is a word and we record the count of corresponding word in that sentence as the value for that dimension. The similarity between the vectors of adjacent sentences in the summary is computed using cosine similarity. The average value of the similarity over all pairs of adjacent sentences is taken as the redundancy score.

We find the average redundancy score for all the summaries produced by each system just as we did with the other verbosity scores in the previous section. We call this score as 'Redundancy score'. Note that the score is based on simple word overlap and cannot capture other subtle redundancy between sentences. In addition, the redundancy is computed only based on words present in adjacent sentences. Repetition and redundancy within the same sentence are also not considered.

7.7.4 Correlations with text quality scores

Next we present the Pearson correlations of the verbosity measures with the gold standard scores in Table 7.16. Since the number of points (systems) is only 22, we indicate whether the correlations are significant at two levels, 0.05 indicated by a '*' superscript and 0.1 indicated by a '.' superscript.

According to true lengths, longer summaries are associated with higher content scores both according to pyramid and content responsiveness evaluations. The length is not significantly related to linguistic quality scores but there is a negative relationship in general. This result conforms to our hypothesis that length alone does not indicate verbosity.

System 14's summary.

Actual length = 254, Predicted length = 292, Verbosity degree = 37

Lawrence at a public inquiry into the teenager's death. Stephen Lawrence, 18, was stabbed in April 1993 at a London bus stop by a group of white attackers. Their identities were not made public. The report, commissioned after police botched the investigation into the 1993 racially motivated murder of a black teenager, Stephen Lawrence, has put pressure on Sir Paul Condon, the Metropolitan Police chief, to resign. LONDON Stephen Lawrence was the picture of a typical English lad. On that April night in 1993, the 18 year old was neatly dressed and well-groomed as he waited for a bus to take him the last leg home to his prim middle class London suburb. Stephen Lawrence, by all accounts a promising student and good-natured young man, was waiting for a bus in suburban Eltham in April 1993 when one or more members of a group of white youths stabbed him to death after using a racial epithet, a witness said. British Home Secretary Jack Straw Tuesday promised new measures to combat racism in the wake of the inquiry into the murder of black teenager Stephen Lawrence. Three of the suspects Neil Acourt, Luke Knight and Gary Dobson were formally acquited of the murder of Stephen three years ago when a private prosecution brought by the Lawrences collapsed. The action against the police comes amid ongoing negotiations with Scotland Yard over compensation. Stephen, 18, was stabbed to death in a racist murder in 1993 at a bus-stop in Eltham, south-east London.

Table 7.14: Summary produced by system 14 for input D0624 shown with the verbosity scores from our model

System 18's summary.

Actual length = 244, Predicted length = 597, Verbosity degree = 353

The government, which has received praise from backers of the Lawrence family for its pursuit of the case, came in for criticism on Monday for actions it took this weekend to prevent publication of a leaked version of the report, which is due to be made public on Wednesday. Sir William Macpherson, a retired High Court justice who was the author of the report and chairman of the eight-month government inquiry, defined institutional racism as 'the collective failure of an organization to provide an appropriate professional service to people because of their color, culture or ethnic origin' reflected, he said, in 'processes, attitudes and behavior which amounts to discrimination through unwitting prejudice, ignorance, thoughtlessness and racist stereotyping.' Richard Norton-Taylor, whose play about Lawrence's killing, 'The Color of Justice,' has been playing to rave reviews in London, said that the attention paid to the Lawrence case and others was a sign that British attitudes toward the overarching authority of the police and other institutions were finally being called into question. She said British authorities and police have learned from the 1993 murder of black teenager Stephen Lawrence by a gang of white youths and the failure of the police to investigate his death adequately. A senior Scotland Yard police officer Wednesday apologized to the parents of a black teenager slain five years ago in a race killing that has become the focus of debate over relations between police and ethnic minorities.

Table 7.15: Summary produced by system 18 for input D0624 shown with the verbosity scores from our model

	Content quality		Linguistic quality		
scores	**Pyramid**	**Cont. resp**	**Non-red**	**Focus**	**Coherence**
actual length	0.64*	0.43*	-0.32	-0.25	-0.32
predicted length	-0.29	-0.11	0.48*	0.39·	0.38·
verbosity degree	-0.47*	-0.23	0.55*	0.44*	0.46*
deviation score	-0.44*	-0.29	0.53*	0.40·	0.42·
redundancy score	-0.01	-0.06	0.06	0.32	0.23

Table 7.16: Pearson correlations between verbosity scores and gold standard summary quality scores. The correlation between actual length of the summary and quality scores is also given in the first row.

Longer summaries on average have better content quality.

On the other hand, when the verbosity scores from our model are examined, all three scores have a negative correlation with content scores. Therefore our verbosity measures appear to be capturing the verbosity arising from the type of content included, different from the length of the summary. The verbosity degree score is the strongest indicator of summary quality with -0.47 correlation (and significant) with pyramid score.

Higher verbosity scores are indicative of lower pyramid scores. At the same time however, verbosity is preferred for linguistic quality. This effect could arise due to the fact these summaries are bags of unordered sentences. Therefore longer sentences and verbose style could be perceived as having greater coherence compared to short and succinct sentences which are jumbled such that it is hard to decipher the full story. It would be interesting to see in future work how these measures are distributed in a corpus of student essays where a reasonable order of sentences and coherence can be assumed.

The plot of the best verbosity score (verbosity degree) and gold standard scores for the significant correlations are shown in Figure 7.3.

The simple redundancy score which we introduced (the last row of the table) does not have any significant relationship to quality scores. One reason could be that most summarization systems make an effort to reduce redundant information [15] and therefore a simple measure of word overlap is not helpful for distinguishing quality. Therefore this result may also be unique to the summarization genre. We need to validate if this

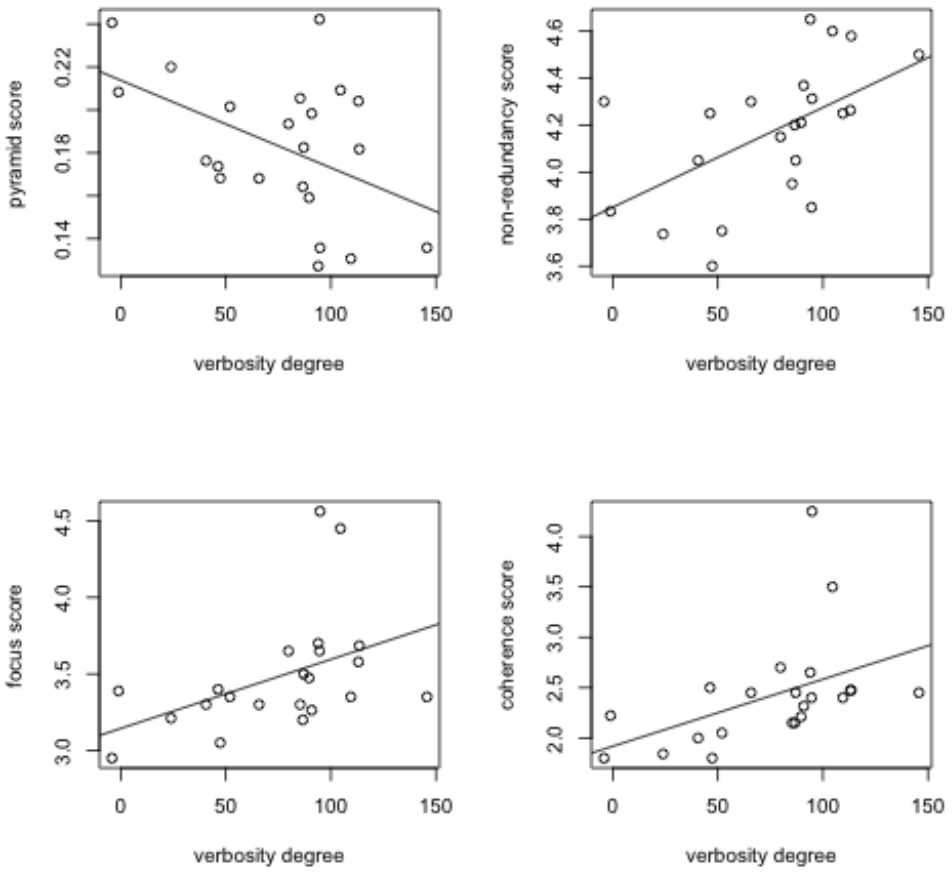

Figure 7.3: Plot of *verbosity degree* measure and gold-standard summary quality scores

aspect is more indicative of quality in other genres of text. In future work, we plan to study in more detail how to implement the redundancy aspect of verbosity with greater sophistication and how we can combine redundancy with our model based on relevance of detail.

7.8 Text quality assessment for science journalism

We now present an experiement using features from our verbosity method for predicting the quality of science news articles. We have assumed in our model that New York Times articles are fairly concise, so it unclear if our model can directly predict verbosity on this set of science articles also taken from the NYT. So we instead add features based on our model to indicate the content type in different parts of the article. We expected that these features could also be useful indicators for quality prediction.

We obtained five snippets of 100 words each randomly chosen from different parts of a test article. For each snippet, we compute the content type features and predict an expected length for the (hypothetical) topic segment containing that snippet. Note that we do not have information about the topic segment containing the snippet. It is only hypothesized. From the predictions, we compute the following features: average predicted length of the snippets, standard deviation of the lengths, minimum and maximum values. These features indicate if different types of content are included in different parts of the article or the type is more or less the same.

A t-test of these scores between the VERY GOOD and TYPICAL categories in our corpus (on a random sample of 1000 articles from each) showed three of these features to vary significantly.

In the VERY GOOD articles, the predicted lengths for the snippets are greater than in the TYPICAL articles. At the same time, there is less deviation in the predicted lengths.

Although significant, when used for classification, these features are not strong enough to perform accurate classification of the categories. They did not provide much accuracies above the baseline. The results of SVM 10-fold classification are shown in Table 7.18. We used the test sets we described in Chapter 3 Section 3.5.

These results suggest that we need to understand better how these features can be

feature	Mean value in VERY GOOD	Mean value in TYPICAL articles
Higher mean values in VERY GOOD articles		
minimum length	201.8	193.9
average length	266.3	261.0
Higher mean values in TYPICAL articles		
standard dev length	108.6	113.0

Table 7.17: Significantly different features from the verbosity model for categories on science news corpus

Any-topic	Same-topic
52.8	51.7

Table 7.18: Accuracies in predicting science news quality using verbosity features

used for such a task.

7.9 Related work

There are no prior studies which have investigated how to predict verbosity. There is however a related notion of idea density which was first discussed by Kintsch and Janice (1973) [74]. They propose idea density as a measure of the number of facts or propositions conveyed in a sentence. Propositions include verbs, adjectives, adverbs, prepositions and conjuctions and the count is normalized by the number of words in the sentence to obtain a score value. Such scores have shown to be indicative of reading difficulty [74] (more propositions indicate more difficult to understand text). On the other hand, a longitudinal study by Snowdon et al. [148] found that low idea density in texts which a person wrote in his early years was highly predictive of low cognitive abilities in later age. Particularly those who developed Alzheimer's disease in later life had scored low on idea density in tests conducted during their early years. A computational tool for measuring idea density was later introduced by Brown et al. (2008) [13]. The tool adjusted the counts for

propositions using a number of rules.

In terms of applications, there is a direct need for controlling the verbosity of a text in tasks such as generation. Users may require to see some or more details and a generation system should be able to adjust the length accordingly. O'Donnell (1997) [117] and Paris et al. (2008) [118] implement ideas about how to adjust content that is produced or displayed depending on space constraints. Their approach involves Rhetorical Structure Theory [97] relations annotated over the full content that is available. Then either relations are given importance values which favor some relations over others, or the idea that satellites of a relation are less important than nuclei is used. Low ranking nodes are dropped when there is a need to shorten the content presented. O'Donnell [117] note that for such a task, apart from content, methods to maintain proper coherence are also important since new paragraph boundaries, reference forms, discourse connectives and punctuation may be required on the shortened text. Our approach has a lot of connections to such work.

It is increasingly acknowledged in automatic summarization as well that systems should include the capacity to generate summaries of different lengths. Kaisser, Hearst and Lowe (2008) [70] present a user study where for different types of search queries, they asked users what summary length would best satisfy a user's need. They found that for different types of queries different length presentations were useful. For example, for general advice, longer summaries were preferred. For queries about people, sometimes, short summaries were preferred and at other times, long and list-like summaries were preferred. Another study related to verbosity in the summarization domain is work by Nenkova, Siddharthan and McKeown (2005) [115]. Here the idea is to predict the familiarity of an entity mentioned in the news for a general audience. Specifically, a distinction is made between hearer-old and hearer-new and discourse-old versus discourse-new entities. When such labels are available for the entities in the summary, a system can use the information to provide descriptions only for unfamiliar entities. Such a strategy would save space by not providing introductions and descriptions for all entities.

Apart from generation and summarization, verbosity models can also be used for quality assessment as we have done in our work. The closest study in this regard is recent work by Agrawal et al. (2011) [1] where the task is to identify chapters of a textbook which

can benefit from the addition of further explanation and details. In this work, they use two clues to determine the enrichment candidates. One factor is syntactic complexity with the idea that more complex material might need to be augmented with extra details. The second factor measures the dispersion of concepts in the text. If the concepts are closely related, then the text is discussing them in detail and on the other hand, the text is vague if the concepts are spread out with weak links between them. A similar study was done by Talukdar and Cohen (2012) [156] where the goal is to predict prerequisite concepts necessary for understanding a new given concept. They run their experiments on Wikipedia articles aiming to predict for a given wiki page, what other pages need to be understood in order to comprehend the current one. This work is less based on the writing of the article but rather on a general idea of what background knowledge is needed. Their features use the structure of Wikipedia links and words which appear in prominent positions of the given article to identify the background concepts. However, their overall goal is also one of text enrichment.

7.10 Future work

This thesis introduced a basic model to relate content type with document verbosity. Many improvements can be made.

Firstly, there is a lot of scope for improving the model for verbosity. Particularly, as we discussed in related work, background knowledge or user's experience is an important factor which will influence perceptions of verbosity and determine which details are necessary and which are irrelvant. For example, in the work by Nenkova, Siddharthan and McKeown (2005) [115], the fact that certain entities in the news are generally familiar to people is employed to shorten the descriptions or reference forms for familiar entities. Similarly, a general background model or one which is specific to a user would also be a necessary component of a verbosity prediction approach. There are several recent attempts at modeling background knowledge which we outlined in the related work above and also some recent work by Peñas and Hovy (2010) [120].

Such an improved model could be useful in some tasks which are currently of great interest. For example, at the DUC conference, the recent summarization evaluations in-

volve an update task[38]. Here the system should assume that the user has read a certain set of documents on a topic. A new set of documents published later on the same topic is now given for summarization. The goal for systems is to only include updates given the background knowledge of the user. We expect that ideas from verbosity modeling can be usefully applied to such tasks.

Another direction for future research is developing ways to actually validate markings of verbosity. While standardly examples of verbose and concise nature appear in writing books, it is unclear how judgements of verbosity can be obtained from people and used for developing a computational method. People are likely to be able to compare two texts for verbosity but less able to identify individual texts as verbose or not. However, a certain amount of supervised material could be quite useful for verbosity prediction. To this end, it may be useful to obtain original and revised drafts of student writing or revisions on community editing websites such as Wikipedia which can be used to create annotations for text verbosity.

7.11 Conclusions

In this chapter we introduced a method to learn the properties of content which are appropriate for articles of different lengths. We used this model trained on concise writing to predict whether a new text is concise or verbose. While we obtained initial success using our method for assessing automatic summaries, there is considerable scope for using the model to do better text quality assessment. Particularly on our corpus of science journalism, features from our verbosity method did not provide better accuracies compared to a random baseline.

[38] http://www.nist.gov/tac/2010/Summarization/index.html

Chapter 8

Discussion

The contributions of this thesis are a framework for text quality prediction and a number of automatic methods to predict different aspects of quality. In this concluding chapter, we reiterate the main ideas in this thesis and lessons we learned from this work.

8.1 Summary of main ideas and results

The central idea in this thesis is the distinction between reader ability and writer skills. In previous work, these two notions were intermingled. In readability for example, suppose a text is proposed as suitable for a fifth grade level student but contains much more complex material than that understandable by an average fifth grade student, then the text is judged to be complex and of lower quality (for that reader). But this judgement sheds little light on the abilities of the writer since the same text could be considered as readable for another audience. In contrast, in text quality, we assume an expert reader as our audience. As a result, whether the content and writing is complex for a reader is not the focus rather whether the text is well-written. This definition of text quality is most suitable for retrieval and recommendation of well-written articles, for providing writing feedback and for use during text generation.

We argued that this setup is similar to educational assessment of writing and we proposed rubrics used in the education field as a suitable framework for defining text quality. This setup identified four core aspects of text quality—conventions, organization,

content and reader interest. Automatic methods for rating these quality dimensions is a good step towards automatically determining the overall quality of a text.

There is considerable progress towards this goal in prior work but there were two main gaps which we aimed to address through our work.

- Automatic measures for text quality were developed mainly for grammar, spelling and organization and other aspects received little focus if any.
- Evaluation of text quality measures was done on examples from news articles, student essays and machine generated text and it is less understood how the measures perform on other genres.

This thesis proposed some solutions to both these problems.

8.1.1 New insights and approaches for predicting text quality

We introduced four methods for text quality prediction which aimed to either focus on hitherto less explored aspects or used new insights to propose measures for already explored quality dimensions.

- **Organization quality based on intentional structure.** We introduced a new syntax based approach for differentiating the organization of well-written articles from incoherent samples. Our idea was based on an assumption that syntax can indicate sentence types and therefore have some relationship to the communicative goal of the sentence. Since intentional structure is well predictive of coherent organization, we supposed that our syntax models would also provide good accuracies in this task. We proposed ways to represent syntax and models that used syntax information to predict organization quality. Our evaluations showed that in the two genres—academic writing and science journalism where we expect similar intentional structure (related to describing a research problem and solution), this approach provides good performance. Further we showed that this approach and predictions are complementary to lexical and coreference statistics for predicting organization quality which were introduced in previous work.

- **Text specificity.** Our measure for specificity tracks general and specific sentences in texts and uses their proportion and sequence for predicting quality. We found that people can annotate these two types of sentences with fair agreement and that a supervised classifier can replicate the annotations with 75% accuracy. Using the automatic classifier, we were able to do large scale evaluation of automatic summaries based on text specificity. We showed that text specificity is indicative of content quality for these texts.

- **Interesting nature of articles.** Our work is one of the first to do a dedicated study of properties of writing and content which can indicate engaging articles. We presented a supervised system to identify interesting science news articles. The features used by the system are quite specialized for the genre of science news and characterise visual nature of writing, narrative structure, beautiful writing and amount of research-related descriptions.

- **Verbosity.** Another as yet unexplored aspect of quality in prior work is verbosity. This thesis presents the first approach to obtain a measurable indicator of text verbosity. We proposed a method to predict verbosity of text based on the compatibility between the type of details included and the length of the article. Wrong type of details indicates either irrelevant details or overly general content both of which are detrimental to quality. We provided a first automatic system for this aspect and showed that the measures we developed are useful for evaluating automatic summaries.

8.1.2 Genre-based study and evaluation of text quality measures

We evaluated our measures on three genres—academic writing, science journalism and automatic summaries. Among these, automatic summaries is the genre that was widely used in prior work, mainly due to the availability of large datasets with manually assigned quality labels. In our work, we have used the automatic summarization genre to evaluate new content related quality measures. We also presented evaluations of our measures on academic and science journalism articles, genres which have been (almost) unexplored so

far for text quality analysis. However, in both genres there is considerable emphasis for good writing and also applications which can greatly benefit from text quality measures.

The benefit of using different genres has been two-fold. Some of the genres inspired and helped us develop new measures for quality. For example, the use of intentional structure-based measure for predicting organization quality was motivated by the previously done wealth of studies on the intentional structure of academic writing. Science journalism on the other hand, showcases research findings in compelling and attractive stories and was suitable for measures related to reader interest. Summarization genre understandably provided support for the development of content quality measures. A second benefit is that we were also able to evaluate any measure we developed on more than one genre. These analyses have provided better understanding of the robustness of our measures across different datasets.

But work on new genres also requires new datasets with quality ratings. Another useful outcome of our work is a new resource for analyzing text quality categories for science journalism. Our corpus contains thousands of articles which were grouped into high and average quality examples using simple heuristics. We hope that this resource will be valuable for other researchers working on text quality tasks. A notable characteristic of this corpus is that the articles groups can be considered as related to overall quality rather than any specific quality aspect. In our work, this property allowed us to study how reader interest measures complement features designed for identifying well-written and easy to read articles.

Below we provide a summary of current performance on this corpus as a reference for future work. We compare features introduced in prior work with those we proposed in this thesis. Each type of measure we introduced was individually evaluated in the respective chapter. Now we provide a view of the performance of all measures in combination with each other. We consider the following classes of features.

Prior work: Features proposed for predicting readability and well-written nature of texts and those for identifying interesting fiction articles. These features correspond to the three categories in Section 6.5.1. This set of features is a comprehensive set representing prior text quality measures for a variety of aspects.

Intentional structure: Features corresponding to different HMM state proportions (after Viterbi decoding on the test article) from the syntax models. We add the combined set of features from the two syntax models, based on productions and d-sequence representations. See Section 4.5 for details.

Text specificity: Features related to proportion of general and specific sentences and specificity of words which we introduced in Section 5.7.

Verbosity: The features which track the content type in different parts of the articles which we used in Section 7.8.

Interesting science: The set of features we specifically proposed for science journalism (Section 6.1).

The evaluation data is the set described in Section 4.5 and we use the same two tasks we have explored so far—differentiating articles from any topic 'any-topic' and those with similar content 'same-topic'. Table 8.1 gives the accuracies of features from prior work and a system that combines prior work with all the new features we introduced in this thesis. (All the above five classes are combined.) We find that for the any-topic setup, the new system after introduction of our measures improves over prior work, with 4% better accuracy for the any-topic setup and 10% for same-topic.

We also show ablation tests for the different classes of features. We see that the generic text quality features from prior work have a huge impact on performance for both tasks. Among our measures, the features related to reader interest and intentional structure have a remarkable impact on the accuracies while the smaller sets of verbosity and specificity features do not make much of a difference in the mix. These results strengthen our claim in this thesis that genre-specific features are also important for text quality prediction.

8.2 Limitations and future work

There are a couple of directions in which the work discussed in this thesis can be improved.

Improvements to models. The accuracy of text quality measures is an area needing much improvement. On the realistic dataset, our science journalism corpus, the highest accuracy obtained is 80%. This performance is quite reasonable. But there is still a

Features	Any-topic	Same-topic
Prior work [P]	73.6	71.8
P + our measures [All]	77.2	81.9
Ablation tests		
All - P	75.7*	77.9*
All - Interesting science	73.3*	78.7*
All - Intentional structure	78.2	78.8*
All - Verbosity	77.1	81.7
All - Specificity	78.7	81.2

Table 8.1: Accuracies for text quality prediction on science journalism articles for different feature sets introduced in this thesis

lot of room for improving the methods. There are a few ways to proceed. Features capturing other aspects of quality can be included. Also, future work should explore how to combine different methods for predicting the overall quality of an article. Another interesting direction is annotation and qualitative studies where people are asked to mark portions of articles which they consider as well-written and attractive. These studies will give us new insights into which properties should be focused upon for developing accurate methods for predicting quality.

Gold standards for evaluation. One of the main challenges for text quality prediction is the non-availability of data with reliable ratings for quality. For reader interest and text specificity aspects, we were able to obtain annotations at least partly and the validation of these measures was easier. On the other hand, for academic writing genre, we followed the easy approach of using permuted articles as negative examples. There is little evidence to show that these examples reflect realistic problems in texts. Perhaps in the context of automatic systems, such texts are likely to be generated by systems and ranking and rating the texts is useful. But these examples are unlikely proxies for problems in texts written by people.

This issue bring up the question of how gold standards should be defined for text

quality prediction. In fact, requiring that we should have manually annotated data for any text quality aspect is an impractical scenario. In fact, people are not likely to be able to rate a specific aspect of quality without being influenced by other aspects. This problem is well-documented in manual evaluation efforts for system summaries. Ratings for different aspects of quality are highly correlated [27, 123]. One tradeoff can be to develop and validate text quality measures for individual aspects using reasonable and well-designed proxies. However, the measures should also then be tested on realistic samples rated for overall quality to understand the benefits of the measures proposed. Overall ratings for quality can be more easily obtained. For example in this thesis, we evaluated our method for organization quality on the academic genre using permutation based examples. Our results there indicated that the models capture properties of good organization versus poorly organized texts. Then we used features from these models to evaluate the quality of science journalism articles, again obtaining reasonable performance. The science journalism corpus was a much more realistic dataset with overall quality ratings. Such evaluations we hope is one direction for future studies.

Use in applications. In this thesis, we evaluated our measures against gold-standard ratings for text quality. But such evaluation is limited in terms of understanding how these measures can be incorporated into the specific applications which motivated the study of these measures. For example, in information retrieval, we may need to balance relevance verus text quality. For writing feedback, there are issues of how quality scores can be translated into specific feedback and estimating when a system can confidently highlight an error or provide feedback. We also proposed scores related to text specificity and verbosity which are quite relevant for automatic summarization. In future work, we plan to investigate how the highly accurate measures from text quality studies can be used in the target applications.

Bibliography

[1] R. Agrawal, S. Gollapudi, A. Kannan, and K. Kenthapadi. Identifying enrichment candidates in textbooks. In *Proceedings of WWW*, pages 483–492, 2011.

[2] G.J. Alred, C.T. Brusaw, and W.E. Oliu. *Handbook of technical writing*. St. Martin's Press, New York, 2003.

[3] N. Asher and A. Lascarides. *Logics of conversation*. Cambridge University Press, 2003.

[4] Y. Attali and J. Burstein. Automated essay scoring with e-rater V.2. *The Journal of Technology, Learning and Assessment*, 4(3), 2006.

[5] C. F. Baker, C. J. Fillmore, and J. B. Lowe. The berkeley framenet project. In *Proceedings of COLING-ACL*, pages 86–90, 1998.

[6] R. Barzilay, N. Elhadad, and K. McKeown. Inferring strategies for sentence ordering in multi-document summar ization. *Journal of Artificial Intelligence Research*, 17:35–55, 2002.

[7] R. Barzilay and M. Lapata. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34, 2008.

[8] R. Barzilay and L. Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of NAACL-HLT*, pages 113–120, 2004.

[9] J. Birke and A. Sarkar. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL*, 2006.

[10] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[11] D. Blum, M. Knudson, and R. M. Henig, editors. *A field guide for science writers: the official guide of the national association of science writers*. Oxford University Press, New York, 2006.

[12] G.H. Bower. Mental imagery and associative learning. In L. Gregg, editor, *Cognition in Learning and Memory*. John Wiley, New York, 1972.

[13] C. Brown, T. Snodgrass, S.J Kemper, R. Herman, and M. Covington. Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, 40:540–545, 2008.

[14] J. Burstein, D. Marcu, and K. Knight. Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *Intelligent Systems, IEEE*, 18(1):32–39, 2003.

[15] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR*, pages 335–336, 1998.

[16] L. Carlson, D. Marcu, and M.E. Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue-Volume 16*, pages 1–10, 2001.

[17] E. Charniak. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of AAAI-IAAI*, pages 598–603, 1997.

[18] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of ACL*, pages 310–318, 1996.

[19] Y.W. Chen and C.J. Lin. Combining svms with various feature selection strategies. *Feature Extraction*, pages 315–324, 2006.

[20] J. Cheung and G. Penn. Utilizing extra-sentential context for parsing. In *Proceedings of EMNLP*, pages 23–33, 2010.

[21] J. Clarke and M. Lapata. Modelling compression with discourse constraints. In *Proceedings of EMNLP-CoNLL*, pages 1–11, 2007.

[22] C. Cocco, R. Pittier, F. Bavaud, and A. Xanthos. Segmentation and clustering of textual sequences: A typological approach. In *Proceedings of RANLP*, pages 427–433, 2011.

[23] M. Collins. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, 1999.

[24] M. Collins and N. Duffy. Convolution kernels for natural language. In *Proceedings of NIPS*, pages 625–632, 2001.

[25] K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag. Personalizing web search results by reading level. In *Proceedings of CIKM*, pages 403–412, 2011.

[26] K. Collins-Thompson and J. Callan. A language modeling approach to predicting reading difficulty. In *Proceedings of HLT-NAACL*, pages 193–200, 2004.

[27] J. M. Conroy and H. T. Dang. Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality. In *Proceedings of COLING*, pages 145–152, 2008.

[28] I. Councill, C. Giles, and M. Kan. ParsCit: An open-source CRF reference string parsing package. In *Proceedings of LREC*, pages 661–667, 2008.

[29] S. Cucerzan and E. Brill. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proceedings of EMNLP*, pages 293–300, 2004.

[30] E. Dale and J. S. Chall. A formula for predicting readability. *Edu. Research Bulletin*, 27(1):11–28, 1948.

[31] R. Dale and A. Kilgarriff. Helping Our Own: Text massaging for computational linguistics as a new shared task. In *Proceedings of INLG*, pages 263–267, 2010.

[32] R. De Felice and S. G. Pulman. A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of COLING*, pages 169–176, 2008.

[33] M. C. De Marneffe, B. MacCartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454, 2006.

[34] P. Diederich. *Measuring Growth in English*. National Council of Teachers of English, 1974.

[35] A. Dimitromanolaki and Ion A. Learning to order facts for discourse planning in natural language generation. In *Proceedings of ENLG*, pages 23–30, 2003.

[36] J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, K. Stratos, K. Yamaguchi, Y. Choi, H. Daume III, A. Berg, and T. Berg. Detecting visual text. In *Proceedings of NAACL-HLT*, pages 762–772, 2012.

[37] A. Dubey, F. Keller, and P. Sturt. Integrating syntactic priming into an incremental probabilistic parser, with an application to psycholinguistic modeling. In *Proceedings of COLING-ACL*, pages 417–424, 2006.

[38] A. Dubey, P. Sturt, and F. Keller. Parallelism in coordination as an instance of syntactic priming: Evidence from corpus-based modeling. In *Proceedings of HLT-EMNLP*, pages 827–834, 2005.

[39] J. Eisenstein and R. Barzilay. Bayesian unsupervised topic segmentation. In *Proceedings of EMNLP*, pages 334–343, 2008.

[40] N. Elhadad. Comprehending technical texts: Predicting and defining unfamiliar terms. In *Proceedings of AMIA*, pages 239–243, 2006.

[41] M. Elsner, J. Austerweil, and E. Charniak. A unified local and global model for discourse coherence. In *Proceedings of NAACL-HLT*, pages 436–443, 2007.

[42] M. Elsner and E. Charniak. Coreference-inspired coherence modeling. In *Proceedings of ACL-HLT: Short Papers*, pages 41–44, 2008.

[43] M. Elsner and E. Charniak. Coreference-inspired coherence modeling. In *Proceedings of ACL-HLT: Short Papers*, pages 41–44, 2008.

[44] M. Elsner and E. Charniak. Disentangling chat with local coherence models. In *Proceedings of ACL-HLT*, pages 1179–1189, 2011.

[45] M. Elsner and E. Charniak. Extending the entity grid with entity-specific features. In *Proceedings of ACL-HLT*, pages 125–129, 2011.

[46] D. Fass. met*: a method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17:49–90, March 1991.

[47] L. Feng, N. Elhadad, and M. Huenerfauth. Cognitively motivated features for readability assessment. In *Proceedings of EACL*, pages 229–237, 2009.

[48] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*, pages 363–370, 2005.

[49] R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32:221 – 233, 1948.

[50] P.W. Foltz, W. Kintsch, and T.K. Landauer. Textual coherence using latent semantic analysis. *Discourse Processes*, 25:285–307, 1998.

[51] P. Fung and G. Ngai. One story, one flow: Hidden markov story models for multilingual multidocument summarization. *ACM Transactions on Speech and Language Processing*, 3(2):1–16, 2006.

[52] M. Galley and K. McKeown. Lexicalized markov grammars for sentence compression. In *Proceedings of HLT-NAACL*, 2007.

[53] M. Gamon, J. Gao, C. Brockett, A. Klementiev, W. B. Dolan, D. Belenko, and L. Vanderwende. Using contextual speller techniques and language modeling for ESL error correction. In *In Proceedings of IJCNLP*, 2008.

[54] D. Graff. The AQUAINT Corpus of English News Text. *Corpus number LDC2002T31, Linguistic Data Consortium, Philadelphia*, 2002.

[55] W. S. Gray and B. E. Leary. *What makes a book readable*. University of Chicago Press, 1935.

[56] B. Grosz, A. Joshi, and S. Weinstein. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226, 1995.

[57] B. Grosz and C. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 3(12):175–204, 1986.

[58] R. Gunning. *The technique of clear writing*. McGraw-Hill; Fouth Printing edition, 1952.

[59] Y. Guo, A. Korhonen, and T. Poibeau. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of EMNLP*, pages 273–283, 2011.

[60] A. Haghighi and D. Klein. Coreference resolution in a modular, entity-centered model. In *Proceedings of HLT-NAACL*, pages 385–393, 2010.

[61] A. Haghighi and L. Vanderwende. Exploring content models for multi-document summarization. In *Proceedings of NAACL-HLT*, pages 362–370, 2009.

[62] M.A.K. Halliday and R. Hasan. *Cohesion in English*. Longman Group Ltd, London, U.K., 1976.

[63] D. Harman and M. Liberman. Tipster complete. *Corpus number LDC93T3A, Linguistic Data Consortium, Philadelphia*, 1993.

[64] B. Howald and M. Abramson. The use of granularity in rhetorical relation prediction. In *Proceedings of *SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, pages 44–48, 2012.

[65] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[66] M. Y. Ivory. *An empirical foundation for automated web interface evaluation*. PhD thesis, University of California, Berkeley, 2001.

[67] M. Y. Ivory and M. A. Hearst. Improving web site design. *IEEE Internet Computing*, 6(2):56–63, 2002.

[68] H. Ji and D. Lin. Gender and animacy knowledge discovery from web-scale n-grams for unsupervised person name detection. In *Proceedings of PACLIC*, 2009.

[69] H. Jing and K. McKeown. Cut and paste based text summarization. In *Proceedings of NAACL*, pages 178–185, 2000.

[70] M. Kaisser, M. A. Hearst, and J. B. Lowe. Improving search results quality by customizing summary lengths. In *Proceedings of ACL-HLT*, pages 701–709, 2008.

[71] N. Karamanis, C. Mellish, M. Poesio, and J. Oberlander. Evaluating centering for information ordering using corpora. *Computational Linguistics*, 35(1):29–46, 2009.

[72] B. Kessler, G. Numberg, and H. Schütze. Automatic detection of text genre. In *Proceedings of ACL-EACL*, pages 32–38, 1997.

[73] J. Y. Kim, K. Collins-Thompson, P. N. Bennett, and S. T. Dumais. Characterizing web content, user interests, and search behavior by reading level and topic. In *Proceedings of WSDM*, pages 213–222, 2012.

[74] W. Kintsch and J. Keenan. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, 5(3):257 – 274, 1973.

[75] D. Klein and C.D. Manning. Accurate unlexicalized parsing. In *Proceedings of ACL*, pages 423–430, 2003.

[76] K. Knight and D. Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1), 2002.

[77] S.M. Kosslyn. *Image and mind*. Harvard University Press, 1980.

[78] M. Krifka, F.J. Pelletier, G.N. Carlson, A. ter Meulen, G. Chierchia, and G. Link. Genericity: an introduction. *The generic book*, pages 1–124, 1995.

[79] M. Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL*, pages 545–552, 2003.

[80] M. Lapata and R. Barzilay. Automatic evaluation of text coherence: Models and representations. In *Proceedings of IJCAI*, pages 1085–1090, 2005.

[81] R. Y. Lau, C. C. Lai, and Y. Li. Mining fuzzy ontology for a web-based granular information retrieval system. In *Proceedings of the 4th International Conference on Rough Sets and Knowledge Technology*, pages 239–246, 2009.

[82] C.W. Leong, S. Hassan, M. E. Ruiz, and R. Mihalcea. Improving query expansion for image retrieval via saliency and picturability. In *Proceedings of the Second international conference on Multilingual and multimodal information access evaluation*, pages 137–142, 2011.

[83] M. Li, Y. Zhang, M. Zhu, and M. Zhou. Exploring distributional similarity based models for query spelling correction. In *Proceedings of COLING-ACL*, pages 1025–1032, 2006.

[84] M. Liakata and L. Soldatova. Guidelines for the annotation of General Scientific Concepts. *JISC Project Report*, 2008.

[85] M. Liakata, S. Teufel, A. Siddharthan, and C. Batchelor. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of LREC*, 2010.

[86] C. Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of Text Summarization Branches Out Workshop, ACL*, pages 74–81, 2004.

[87] C. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL*, pages 1085–1090, 2003.

[88] D. Lin, K. W. Church, H. Ji, S. Sekine, D. Yarowsky, S. Bergsma, K. Patil, E. Pitler, R. Lathbury, V. Rao, K. Dalwani, and S. Narsale. New tools for web-scale n-grams. In *Proceedings of LREC*, 2010.

[89] Z. Lin, M. Kan, and H. Ng. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of EMNLP*, pages 343–351, 2009.

[90] Z. Lin, H. Ng, and M. Kan. Automatically evaluating text coherence using discourse relations. In *Proceedings of ACL-HLT*, pages 997–1006, 2011.

[91] A. Louis, A. Joshi, and A. Nenkova. Discourse indicators for content selection in summarization. In *Proceedings of SIGDIAL*, pages 147–156, 2010.

[92] A. Louis, A. Joshi, R. Prasad, and A. Nenkova. Using entity features to classify implicit discourse relations. In *Proceedings of SIGDIAL*, pages 59–62, 2010.

[93] A. Louis and A. Nenkova. Creating local coherence: An empirical assessment. In *Proceedings of HLT-NAACL*, pages 313–316, 2010.

[94] A. Louis and A. Nenkova. Automatically assessing machine summary content without a gold-standard. *Computational Linguistics*, 2012.

[95] A. Louis and A. Nenkova. A coherence model based on syntactic patterns. In *Proceedings of EMNLP-CoNLL*, pages 1157–1168, 2012.

[96] Y. Ma, E. Fosler-Lussier, and R. Lofthus. Ranking-based readability assessment for early primary children's literature. In *Proceedings of NAACL-HLT*, pages 548–552, 2012.

[97] W.C. Mann and S.A. Thompson. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8, 1988.

[98] D. Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press, 2000.

[99] D. Marcu and A. Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of ACL*, pages 368–375, 2001.

[100] M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1994.

[101] R. Mason and E. Charniak. Extractive multi-document summaries should explicitly not contain document-specific content. In *Proceedings of the Workshop on Au-*

tomatic Summarization for Different Genres, Media, and Languages, ACL-HLT, pages 49–54, 2011.

[102] T. Mathew and G. Katz. Supervised categorization for habitual versus episodic sentences. In *Sixth Midwest Computational Linguistics Colloquium. Indiana University Bloomington, May*, pages 2–3, 2009.

[103] R. McDonald. Discriminative sentence compression with soft syntactic evidence. In *Proceedings of EACL*, 2006.

[104] N. McIntyre and M. Lapata. Learning to tell tales: A data-driven approach to story generation. In *Proceedings of ACL-IJCNLP*, pages 217–225, 2009.

[105] Q. Mei and C. Zhai. Generating impact-based summaries for scientific literature. In *Proceedings of ACL-HLT*, pages 816–824, 2008.

[106] R. Mihalcea and C. Strapparava. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, 22(2):126–142, 2006.

[107] E. Miltsakaki and A. Troutt. Read-X: Automatic evaluation of reading difficulty of web text. In *Proceedings of E-Learn*, 2007.

[108] R. J. Mooney and L. Roy. Content-based book recommending using learning for text categorization. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 195–204, 2000.

[109] R. Mulkar-Mehta, J. Hobbs, and E. Hovy. Granularity in natural language discourse. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 360–364, 2011.

[110] D.M. Murray. *Learning by Teaching: Selected Articles on Writing and Teaching*. Boynton/Cook Publishers Inc., 1982.

[111] A. Nakhimovsky. Aspect, aspectual class, and the temporal structure of narrative. *Computational Linguistics*, 14(2):29–43, June 1988.

[112] A. Nenkova. Entity-driven rewrite for multi-document summarization. In *Proceedings of IJCNLP*, 2008.

[113] A. Nenkova and A. Louis. Can you summarize this? identifying correlates of input difficulty for multi-document summarization. In *Proceedings of ACL-HLT*, pages 825–833, 2008.

[114] A. Nenkova, R. Passonneau, and K. McKeown. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4(2):4, 2007.

[115] A. Nenkova, A. Siddharthan, and K. McKeown. Automatically learning cognitive status for multi-document summarization of newswire. In *Proceedings of HLT-EMNLP*, pages 241–248, 2005.

[116] M.C. Nisbet, D. Brossard, and A. Kroepsch. Framing science: The stem cell controversy in an age of press/politics. *The International Journal of Press/Politics*, 8(2):36–70, 2003.

[117] M. O'Donnell. Variable-length on-line document generation. In *Proceedings of the 6th European Workshop on Natural Language Generation*, 1997.

[118] C. Paris, N. Colineau, A. Lampert, and J. Giralt Duran. Generation under space constraints. In *Proceedings of COLING*, pages 127–129, 2008.

[119] M. Pazzani, J. Muramatsu, and D. Billsus. Syskill & Webert: Identifying interesting web sites. In *Proceedings of AAAI*, pages 54–61, 1996.

[120] A. Peñas and E. Hovy. Semantic enrichment of text with background knowledge. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 15–23, 2010.

[121] P. Petrenz and B. Webber. Stable classification of text genres. *Computational Linguistics*, 37(2):385–393, 2011.

[122] E. Pitler, A. Louis, and A. Nenkova. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of ACL-IJCNLP*, pages 683–691, 2009.

[123] E. Pitler, A. Louis, and A. Nenkova. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of ACL*, 2010.

[124] E. Pitler and A. Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of EMNLP*, pages 186–195, 2008.

[125] E. Pitler and A. Nenkova. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of ACL-IJCNLP*, pages 13–16, 2009.

[126] H. Po. News and its communicative quality: the inverted pyramid: when and why did it appear? *Journalism Studies*, 4(4):501–511, 2003.

[127] M. Poesio, R. Stevenson, B. Di Eugenio, and J. Hitzeman. Centering: A parametric theory and its instantiations. *Computational Linguistics*, pages 309–363, 2004.

[128] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. The Penn Discourse Treebank 2.0. In *Proceedings of LREC*, 2008.

[129] A. C. Purves. Reflections on research and assessment in written composition. *Research in the Teaching of English*, 26(1):108–122, 1992.

[130] V. Qazvinian, D. Radev, and A. Ozgur. Citation summarization through keyphrase extraction. In *Proceedings of COLING*, pages 895–903, 2010.

[131] V. Qazvinian and D. R. Radev. Scientific paper summarization using citation summary networks. In *Proceedings of COLING*, pages 689–696, 2008.

[132] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2011.

[133] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257 –286, 1989.

[134] D. Radev, M. Joseph, B. Gibson, and P. Muthukrishnan. A bibliometric and network analysis of the field of computational linguistics. *Journal of the American Society for Information Science and Technology*, 2009.

[135] D. Radev, P. Muthukrishnan, and V. Qazvinian. The ACL anthology network corpus. In *Proceedings of ACL Workshop on Natural Language Processing and Information Retrieval for Digital Libraries*, 2009.

[136] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning. A multi-pass sieve for coreference resolution. In *Proceedings of EMNLP*, pages 492–501, 2010.

[137] D. Ramage, D. Hall, R. Nallapati, and C.D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of EMNLP*, pages 248–256, 2009.

[138] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[139] N. Reiter and A. Frank. Identifying generic noun phrases. In *Proceedings of ACL*, pages 40–49, 2010.

[140] D. Reitter, J. Moore, and F. Keller. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 685–690, 2006.

[141] J. Reynar and A. Ratnaparkhi. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 16–19, 1997.

[142] E. Sandhaus. The New York Times Annotated Corpus. *Corpus number LDC2008T19, Linguistic Data Consortium, Philadelphia*, 2008.

[143] B. Schiffman. *Learning to Identify New Information*. PhD thesis, Columbia University, 2005.

[144] S. Schwarm and M. Ostendorf. Reading level assessment using support vector machines and statistical language models. In *Proceedings of ACL*, pages 523–530, 2005.

[145] H.A Semetko and P.M Valkenburg. Framing european politics: a content analysis of press and television news. *Journal of Communication*, 50(2):93–109, 2000.

[146] E. Shutova, L. Sun, and A. Korhonen. Metaphor identification using verb and noun clustering. In *Proceedings of COLING*, pages 1002–1010, 2010.

[147] L. Si and J. Callan. A statistical model for scientific readability. In *Proceedings of CIKM*, pages 574–576, 2001.

[148] D.A. Snowdon, S.J. Kemper, J.A. Mortimer, L.H. Greiner, D.R. Wekstein, and W.R. Markesbery. Linguistic ability in early life and cognitive function and alzheimer's disease in late life. *Jama*, 275(7):528–532, 1996.

[149] R. Soricut and D. Marcu. Discourse generation using utility-trained coherence models. In *Proceedings of COLING-ACL*, pages 803–810, 2006.

[150] V. Spandel. *Creating Writers Through 6-Trait Writing: Assessment and Instruction*. Allyn and Bacon, Inc., 2004.

[151] C. Sporleder and A. Lascarides. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14:369–416, 2008.

[152] S. H. Stocking. *The New York Times Reader: Science and Technology*. CQ Press, Washington DC, 2010.

[153] P.J. Stone, J. Kirsh, and Cambridge Computer Associates. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, 1966.

[154] J. Swales. *Genre analysis: English in academic and research settings*, volume 11. Cambridge University Press, 1990.

[155] J. Swales and C. Feak. *Academic writing for graduate students: A course for non-native speakers of English*. Ann Arbor: University of Michigan Press, 1994.

[156] P. Talukdar and W. Cohen. Crowdsourced comprehension: predicting prerequisite structure in wikipedia. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 307–315, 2012.

[157] J. Tetreault, J. Foster, and M. Chodorow. Using parse features for preposition selection and error detection. In *Proceedings of ACL*, pages 353–358, 2010.

[158] S. Teufel. *Argumentative Zoning: Information Extraction from Scientific Text*. PhD thesis, University of Edinburgh, 2000.

[159] S. Teufel. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. Center for the Study of Language and Information - Lecture Notes Series, 2010.

[160] S. Teufel, J. Carletta, and M. Moens. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of EACL*, pages 110–117, 1999.

[161] S. Teufel and M. Kan. Robust argumentative zoning for sensemaking in scholarly documents. In *Proceedings of the 2009 international conference on Advanced language technologies for digital libraries*, pages 154–170, 2011.

[162] S. Teufel and M. Moens. What's yours and what's mine: determining intellectual attribution in scientific text. In *Proceedings of EMNLP*, pages 9–17, 2000.

[163] S. Teufel and M. Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.

[164] S. Teufel, A. Siddharthan, and D. Tidhar. Automatic classification of citation function. In *Proceedings of EMNLP*, pages 103–110, 2006.

[165] O. Uryupina. High-precision identification of discourse new and unique noun phrases. In *Proceedings of ACL*, pages 80–86, 2003.

[166] P.M. Valkenburg, H.A. Semetko, and C.H. De Vreese. The effects of news frames on readers' thoughts and recall. *Communication research*, 26(5):550–569, 1999.

[167] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of CHI*, pages 319–326, 2004.

[168] R. L. Weide. The cmu pronunciation dictionary, release 0.6. *http://www.speech.cs.cmu.edu/cgi- bin/cmudict*, 1998.

[169] E. Weitkamp. British newspapers privilege health and medicine topics over other science news. *Public Relations Review*, 29(3):321–333, 2003.

[170] J. M. Williams. *Style: Toward Clarity and Grace*. The University of Chicago Press, 1990.

[171] M. Wilson. MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods*, 20(1):6–10, 1988.

[172] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT-EMNLP*, pages 347–354, 2005.

[173] J. Zhao and M. Kan. Domain-specific iterative readability computation. In *Proceedings of JDCL*, pages 205–214, 2010.

[174] Y. Zhao, G. Karypis, and U. Fayyad. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10:141–168, 2005.