Revisiting Readability: A Unified Framework for Predicting Text Quality

Emily Pitler
Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
epitler@seas.upenn.edu

Ani Nenkova
Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
nenkova@seas.upenn.edu

Abstract

We combine lexical, syntactic, and discourse features to produce a highly predictive model of human readers' judgments of text readability. This is the first study to take into account such a variety of linguistic factors and the first to empirically demonstrate that discourse relations are strongly associated with the perceived quality of text. We show that various surface metrics generally expected to be related to readability are not very good predictors of readability judgments in our Wall Street Journal corpus. We also establish that readability predictors behave differently depending on the task: predicting text readability or ranking the readability. Our experiments indicate that discourse relations are the one class of features that exhibits robustness across these two tasks.

1 Introduction

The quest for a precise definition of text quality—pinpointing the factors that make text flow and easy to read—has a long history and tradition. Way back in 1944 Robert Gunning Associates was set up, offering newspapers, magazines and business firms consultations on clear writing (Gunning, 1952). In education, teaching good writing technique and grading student writing has always been of key importance (Spandel, 2004; Attali and Burstein, 2006). Linguists have also studied various aspects of text flow, with cohesion-building devices in English (Halliday and Hasan, 1976), rhetorical structure theory (Mann and Thompson, 1988) and centering theory (Grosz et al., 1995) among the most influential contributions.

Still, we do not have unified computational models that capture the interplay between various aspects of readability. Most studies focus on a single factor contributing to readability for a given intended audience. The use of rare words or technical terminology for example can make text difficult to read for certain audience types (Collins-Thompson and Callan, 2004; Schwarm and Ostendorf, 2005; Elhadad and Sutaria, 2007). Syntactic complexity is associated with delayed processing time in understanding (Gibson, 1998) and is another factor that can decrease readability. Text organization (discourse structure), topic development (entity coherence) and the form of referring expressions also determine readability. But we know little about the relative importance of each factor and how they combine in determining perceived text quality.

In our work we use texts from the Wall Street Journal intended for an *educated adult audience* to analyze readability factors including vocabulary, syntax, cohesion, entity coherence and discourse. We study the association between these features and reader assigned readability ratings, showing that discourse and vocabulary are the factors most strongly linked to text quality. In the easier task of text quality ranking, entity coherence and syntax features also become significant and the combination of features allows for ranking prediction accuracy of 88%. Our study is novel in the use of gold-standard discourse features for predicting readability and the simultaneous analysis of various readability factors.

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 186–195,
Honolulu, October 2008.

2 Related work

2.1 Readability with respect to intended readers

The definition of what one might consider to be a well-written and readable text heavily depends on the intended audience (Schriver, 1989). Obviously, even a superbly written scientific paper will not be perceived as very readable by a lay person and a great novel might not be appreciated by a third grader. As a result, the vast majority of prior work on readability deals with labeling texts with the appropriate school grade level. A key observation in even the oldest work in this area is that the vocabulary used in a text largely determines its readability. More common words are easier, so some metrics measured text readability by the percentage of words that were not among the N most frequent in the language. It was also observed that frequently occurring words are often short, so word length was used to approximate readability more robustly than using a predefined word frequency list. Standard indices were developed based on the link between word frequency/length and readability, such as Flesch-Kincaid (Kincaid, 1975), Automated Readability Index (Kincaid, 1975), Gunning Fog (Gunning, 1952), SMOG (McLaughlin, 1969), and Coleman-Liau (Coleman and Liau, 1975). They use only a few simple factors that are designed to be easy to calculate and are rough approximations to the linguistic factors that determine readability. For example, Flesch-Kincaid uses the average number of syllables per word to approximate vocabulary difficulty and the average number of words per sentence to approximate syntactic difficulty.

In recent work, the idea of linking word frequency and text readability has been explored for making medical information more accessible to the general public. (Elhadad and Sutaria, 2007) classified words in medical texts as familiar or unfamiliar to a general audience based on their frequencies in corpora. When a description of the unfamiliar terms was provided, the perceived readability of the texts almost doubled.

A more general and principled approach to using vocabulary information for readability decisions has been the use of language models. For any given text, it is easy to compute its likelihood under a given language model, i.e. one for text meant for children, or for text meant for adults, or for a given grade level. (Si and Callan, 2001), (Collins-Thompson and Callan, 2004), (Schwarm and Ostendorf, 2005), and (Heilman et al., 2007) used language models to predict the suitability of texts for a given school grade level. But even for this type of task other factors besides vocabulary use are at play in determining readability. Syntactic complexity is an obvious factor: indeed (Heilman et al., 2007) and (Schwarm and Ostendorf, 2005) also used syntactic features, such as parse tree height or the number of passive sentences, to predict reading grade levels. For the task of deciding whether a text is written for an adult or child reader, (Barzilay and Lapata, 2008) found that adding entity coherence to (Schwarm and Ostendorf, 2005)'s list of features improves classification accuracy by 10%.

2.2 Readability as coherence for competent language users

In linguistics and natural language processing, the text properties rather than those of the reader are emphasized. Text coherence is defined as the ease with which a person (tacitly assumed to be a competent language user) understands a text. Coherent text is characterized by various types of cohesive links that facilitate text comprehension (Halliday and Hasan, 1976).

In recent work, considerable attention has been devoted to entity coherence in text quality, especially in relation to information ordering. In many applications such as text generation and summarization, systems need to decide the order in which selected sentences or generated clauses should be presented to the user. Most models attempting to capture local coherence between sentences were based on or inspired by centering theory (Grosz et al., 1995), which postulated strong links between the center of attention in comprehension of adjacent sentences and syntactic position and form of reference. In a detailed study of information ordering in three very different corpora, (Karamanis et al., to appear) assessed the performance of various formulations of centering. Their results were somewhat unexpected, showing that while centering transition preferences were useful, the most successful strategy for information ordering was based on avoid-

ing rough shifts, that is, sequences of sentences that share no entities in common. This supports previous findings that such types of transitions are associated with poorly written text and can be used to improve the accuracy of automatic grading of essays based on various non-discourse features (Miltsakaki and Kukich, 2000). In a more powerful generalization of centering, Barzilay and Lapata (2008) developed a novel approach which doesn't postulate a preference for any type of transition but rather computes a set of features that capture transitions of all kinds in the text and their relative proportion. Their entity coherence features prove to be very suitable for various tasks, notably for information ordering and reading difficulty level.

Form of reference is also important in well-written text and appropriate choices lead to improved readability. Use of pronouns for reference to highly salient entities is perceived as more desirable than the use of definite noun phrases (Gordon et al., 1993; Krahmer and Theune, 2002). The syntactic forms of first mention—when an entity is first introduced in a text—differ from those of subsequent mentions (Poesio and Vieira, 1998; Nenkova and McKeown, 2003) and can be exploited for improving and predicting text coherence (Siddharthan, 2003; Nenkova and McKeown, 2003; Elsner and Charniak, 2008).

3 Data

The objective of our study is to analyze various readability factors, including discourse relations, because few empirical studies exist that directly link discourse structure with text quality. In the past, subsections of the Penn Treebank (Marcus et al., 1994) have been annotated for discourse relations (Carlson et al., 2001; Wolf and Gibson, 2005). For our study we chose to work with the newly released Penn Discourse Treebank which is the largest annotated resource which focuses exclusively on implicit local relations between adjacent sentences and explicit discourse connectives.

3.1 Discourse annotation

The Penn Discourse Treebank (Prasad et al., 2008) is a new resource with annotations of discourse connectives and their senses in the Wall Street Journal portion of the Penn Treebank (Marcus et al., 1994). All *explicit* relations (those marked with a discourse connective) are annotated. In addition, each adjacent pair of sentences within a paragraph is annotated. If there is a discourse relation, then it is marked *implicit* and annotated with one or more connectives. If there is a relation between the sentences but adding a connective would be inappropriate, it is marked *AltLex*. If the consecutive sentences are only related by entity-based coherence (Knott et al., 2001) they are annotated with *EntRel*. Otherwise, they are annotated with *NoRel*.

Besides labeling the connective, the PDTB also annotates the *sense* of each relation. The relations are organized into a hierarchy. The top level relations are Expansion, Comparison, Contingency, and Temporal. Briefly, an expansion relation means that the second clause continues the theme of the first clause, a comparison relation indicates that something in the two clauses is being compared, contingency means that there is a causal relation between the clauses, and temporal means they occur either at the same time or sequentially.

3.2 Readability ratings

We randomly selected thirty articles from the Wall Street Journal corpus that was used in both the Penn Treebank and the Penn Discourse Treebank.[1] Each article was read by at least three college students, each of whom was given unlimited time to read the texts and perform the ratings.[2] Subjects were asked the following questions:

- How well-written is this article?
- How well does the text fit together?
- How easy was it to understand?
- How interesting is this article?

For each question, they provided a rating between 1 and 5, with 5 being the best and 1 being the worst.

[1] One of the selected articles was missing from the Penn Treebank. Thus, results that do not require syntactic information (Tables 1, 2, 4, and 6) are over all thirty articles, while Tables 3, 5, and 7 report results for the twenty-nine articles with Treebank parse trees.

[2] (Lapata, 2006) found that human ratings are significantly correlated with self-paced reading times, a more direct measure of processing effort which we plan to explore in future work.