Automatic Readability Classification of Crowd-Sourced Data based on Linguistic and Information-Theoretic Features

Zahurul Islam and Alexander Mehler

AG Texttechnology, Institut für Informatik,
Goethe-Universität, Frankfurt, Germany
zahurul,mehler@em.uni-frankfurt.de

Abstract. This paper presents a classifier of text readability based on information-theoretic features. The classifier was developed based on a linguistic approach to readability that explores lexical, syntactic and semantic features. For this evaluation we extracted a corpus of 645 articles from Wikipedia together with their quality judgments. We show that information-theoretic features perform as well as their linguistic counterparts even if we explore several linguistic levels at once.

Keywords. Text readability, Wikipedia, enthropy, information transmission, evaluation of features.

Clasificación automática de la legibilidad de datos de fuentes múltiples basada en características lingüísticas y de la teoría de información

Resumen. En este trabajo se presenta un clasificador de la legibilidad de textos basado en las características de la teoría de información. El clasificador ha sido desarrollado en base del enfoque lingüístico a la legibilidad usando las características léxicas, sintácticas y semánticas. Para esta evaluación se extrajo un corpus de 645 artículos de Wikipedia, junto con sus evaluaciones de calidad. Se demuestra que las características mencionadas tienen buen desempeño, incluso en el caso cuando se exploran varios niveles lingüísticos a la vez.

Palabras clave. Legibilidad de textos, Wikipedia, entropía, transmisión de información, evaluación de características.

1 Introduction

The readability of a text relates to how well and how easily it conveys its meaning to its readers. There are many text-related factors that influence readability. They range from simple features such as type face, font size or text vocabulary to complex features such as syntactic, semantic, rhetorical, or genre structure.

Many professionals, such as teachers, journalists, or editors, create text for their audiences and routinely check its readability. With our classifier, we explore the task of automatically classifying documents according to different readability levels. As input, this function operates on various statistics of lexical, syntactic, semantic and other text features.

Automatic readability classification can be useful for many Natural Language Processing (NLP) applications. Automatic essay grading can benefit from readability classification as a guide to how good an essay actually is. Similarly, a search engine can use a readability classifier to rank its generated search results. Automatically generated documents, for example documents generated by text summarization systems or machine translation systems, tend to be error-prone and not very readable. In this case, a readability classification system can be used to filter out documents that are less readable. The system can also be used to evaluate machine translation output.

In this paper, we provide an information-theoretic approach to readability classifiers that uses a crowd-sourced corpus of readability assessments. More specifically, we build a corpus of texts together with readability assessments that were extracted from Wikipedia. Wikipedia is a product of collaborative crowd-sourcing. Articles in Wikipedia differ with respect to their quality in terms of readability. Many of them are well written, while for others, the authors have not even reached an agreement among themselves regarding the topic. Since 2010, Wikipedia offers an *article feedback tool* that allows readers to assess the quality of articles. Contributors can give their opinion about the feedback in terms of quality. The readability measure is reflected by the *Well-written* feature of the feedback tool, which places a document in a class (one to five stars: incomprehensible,

difficult to understand, adequate quality, good clarity, exceptional clarity).

According to Wikipedia[1], around 40,000 ratings are submitted everyday, 97% of them by anonymous users. 90% of users claim that the page ratings are useful. It should be noted that the *Well-written* score might be affected by contributors' perception of the non-linguistic aspects of the article, such as the objectivity or bias of the article. But, as a corrective, the article ratings are evaluated by a group of experts of the corresponding subject areas. A recent study by [17] showed that the accuracy of many of the articles in Wikipedia is comparable to Wikipedia's rival, the Encyclopedia Britannica.

Our readability classifier is based on lexical, linguistic and information-theoretic features. We evaluate this classifier in comparison with a linguistic approach to readability that explores lexical, syntactic and semantic features. Some of the linguistic features presented in this work are being used here for the first time for readability classification. Our evaluation shows that simple information-theoretic features perform equally well in comparison to their linguistic counterparts, even if they explore lexical, syntactic and semantic text features at once.

The paper is organized as follows: Section 2 discusses related work followed by an introduction of our corpus in Section 3. The features used for classification are described in Section 4, and our experiment and evaluation in Section 5. Our experiments and findings are discussed in Section 6. Finally, we present conclusions and future work in Section 7.

2 Related Work

At present, there is no standard approach to measuring text quality in terms of readability. According to [32], a readable article should contain sentences averaging between 14 to 22 words. If the average length of sentences is longer than 22 words, then the article can lose clarity, while if the average sentence length is less than 14 words, then the ideas can appear discontinuous [32].

In the early stages of readability research, fairly simple features were used due to a lack of linguistic resources and computational power. *Average Sentence length* (ASL) is one of them. [5, 6] showed that reading difficulty is a linear function of the ASL, combined with the percentage of rare words. They listed $3,000$ commonly known words for the 4^{th} grade (i.e. children approximately 10 years old). Their assumption was that a text containing more familiar words will be easier to read and understand.

[18] also considered the numbers of sentences and complex words in order to measure text readability. This research focused on word complexity that is directly related to the number of syllables in the word: the smaller number of syllables, the easier the word. The Flesch-Kincaid readability index [26] considers the average number of words per sentence and the average number of syllables per word. They proposed two different formulas, one for measuring how easy a text is to read and the other one for measuring grade level. [39] also designed a readability index for the *US Air force* that uses the average number of characters in a word and the average number of words in a sentence. [33] and [9] show that these methods have significant drawbacks. Longer sentences are not always syntactically complex and counting the number of syllables of a single word does not show word difficulty. With recent advancements of NLP tools, a new class of linguistically-motivated text features is now available.

[38, 34, 2, 24, 8] use statistical language models to classify texts for their readability. The motivation of using a language model is that a probabilistic model provides a prediction of how likely a given sentence can be generated by the same underlying process that generated a corpus of texts of different readability classes. They show that trigrams are more informative than bigram and unigram models.

Parts of Speech (POS)-based grammatical features were shown to be useful in readability classification [9, 2, 10]. The number of common nouns give an approximation of the number of entities in a text that the reader has to keep in memory to understand the text. POS based features outperform language-model-based and syntactic features in [10].

Text readability is affected by syntactic constructions [34, 3, 20, 21, 30]. In this line of research, [3] show, for example, that multiple noun phrases in a single sentence require the reader to remember more items. Multiple verb phrases in a sentence may indicate the presence

[1] http://en.wikipedia.org/wiki/Wikipedia:Article_Feedback_Tool/Version_5

of explicit discourse relation in the sentence. Piter and Nenkova [34] showed the strongest correlation between text readability and the number of verb phrases.

On the semantic level, a paragraph that refers to many entities at once burdens the reader since she has to keep track of these entities, their semantic representations and how these entities are related. Texts that refer to many entities are extremely difficult to understand for people with intellectual disabilities [9].

Researchers also experimented with semantic features like *lexical chains*, *discourse relations* and *entity grids* [10, 3]. It has been shown that these features are useful for readability classification.

[23] used *entropy* and *Kullback-Leibler divergence-based features* to classify articles in text books. These information-theoretic features give a 50% rise of accuracy and F-score over the baseline system that uses three traditional readability formulas. They have shown that these features are very useful for readability classification, especially for low-resource languages.

Some of the features used in our work are also used in some of the approaches being described above. These features can not be directly compared, however, because different data sets are used. Also different tools are used to extract various linguistic properties.

3 Corpus

In this section, we describe the crowd-sourced corpus that we extracted from Wikipedia and that we used for evaluating information-theoretic features in comparison to linguistic features.

The Wikimedia foundation provides feedback data as *raw rating data* and *article summary data*. The *raw rating data* shows the average raw ratings of an article and the *article summary data* shows the *total rating score* and the *number of ratings submitted*. The *article summary data* is taken into account in the text extraction process, because it helps us to put a threshold on the process of text extraction. Article length is a factor when extracting text features for classification. The rating of an article that is evaluated by many readers is more reliable than the rating of an article that is evaluated by a small number of readers. These two factors were taken into account in the selection of articles for text extraction from Wikipedia: all extracted articles are rated by at least 10 readers and contain at least 10 sentences. For our experiment, we used the *article summary data* of September 19, 2011[2] and a Wikipedia dump from August 2011.

Table 1. Statistics of the Wikipedia-based readability corpus

Classes	Articles	#Tokens	#Types	Text Length
one *and* two	61	115,729	24,934	36.40
three	169	869,607	95,578	81.98
four	209	2,402,704	184,357	164.50
five	206	2,561,115	201,682	175.26

To extract Wikipedia articles, we utilized a freely available extraction tool[3]. From the *article summary data*, five classes of Wikipedia articles were extracted. Each class corresponds to a readability level (e.g., one, two, three, four and five) in ascending order of readability. The class of an article is determined by the *mean rating score*. As a result of our extraction, class *one* contained 12 and class *two* contained 49 articles. To avoid problems of data sparseness these two classes were merged. Table 1 shows the final corpus that we used for our experiment.

4 Features

In this paper, we compare an information-theoretic classifier of text readability with a classifier based on linguistic features. We start by describing the classifier based on linguistic features.

4.1 Linguistic Features

The literature explores a variety of linguistic indicators of readability. We developed a classifier of text readability based on lexical, syntactic and semantic features. We first describe lexical features used by this classifier.

[2] http://dumps.wikimedia.org/other/articlefeedback/aap_combined-20110919.csv.gz

[3] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

4.1.1 Lexical Features

This includes the *avg. sentence length*, the *avg. number of syllables* and the *avg. number of difficult words* (more than 8 letters). The *Average Sentence Length* is a quantitative measure of syntactic complexity. The *Average Word Length* is another lexical feature that is useful for readability classification. For example: the word *biodegradable* will be harder to pronounce, spell and understand for some readers.

The *type token ratio* (TTR), which indicates the lexical density of text, has been considered as a readability feature. Low lexical densities involve a great deal of repetition with the same words occurring again and again. Conversely, high lexical density shows the diverseness of a text. Our assumption is that a diverse text is supposed to be difficult for readers, generally resulting in lower readability classes. The TTR will consider *Book* and *books* as different words and word types even though these are lexically the same. To circumvent this deficit, we compute the TTR on the level of types as lemmata. All of these features are computed at the sentence level and on the level of the text as a whole. Ten lexical features are listed in Table 2.

4.1.2 POS-based Features

Some of the previous research pointed out that the difficulty level of a text is influenced by the number of POS it contains. For example, the number of named entities in a text can be approximated by the number of words tagged as *noun*. Readers need to co-refer pronouns with appropriate nouns, which might impose an extra burden for less-skilled readers. The number of definite articles provide a measurement of how abstract a text is since an abstract text will have fewer definite articles. In this paper, we focus on seven POS categories (N, V, PRO, ADJ, ADV, PREP and DET) where for each category X two features are computed: the *avg. number of X in a sentence* and the *number of X in a text*.

4.1.3 Syntactic Features

A text can be less readable due to unusual linguistic constructions. As an example, [3] found that an article that is written for adult readers contains more noun phrases than an article that is written for children. Multiple verb phrases in a sentence may also play a role in making a sentence difficult to read.

Subordinate clauses, according to [34], correlated positively with text readability. Therefore, a sentence with a subordinate clause will be difficult for children or a less-skilled reader. In our experiment, we focus on *VP*s, *NP*s, *PP*s, *subordinate clauses* and *embedded clauses* (i.e., clauses in argument position). For each of these syntactic classes *XP*, we compute two features: the *avg. number of XP in a sentence* and the *number of XP in a text*. This results in 10 syntactic features. Note that we used the Stanford PCFG parser [27] for parsing Wikipedia articles.

4.1.4 Semantic Features

The semantics of a text plays an important role in assessing its readability. A number of semantic indicators of readability, which are not accounted for in related studies, are *co-reference* [22], *frame semantics* [11, 1] and *semantic roles* [16]. The reachability of antecedents of anaphoric expressions is subject to the *Right Frontier Constraint* [37]. That is, anaphoric expressions can only be attached to elements that lie on the right hand side of the text tree or graph spanned by rhetorical relations [31]. From this perspective, it should be easier to resolve anaphora that are close to their antecedents in terms of their distances in the text structure graph as spanned, for example, by rhetorical relations [31]. Thus, the longer the distance between an anaphoric expression and its antecedent, the less readable the text. In order to explore this feature, we use the tool named *Reconcile* [41] to annotate Wikipedia articles with co-reference information.

A semantic frame is a coherent structure of related concepts [1]. A semantic frame contains many facts that represent characteristic features, attributes and functions of a referent. As the context has to be considered during the reading of such sentences, a sentence with a semantic frame will be harder to read for less-skilled readers. Further, semantic role labeling is a task of shallow semantic parsing where each predicate is mapped onto its semantic roles. Our hypothesis is that the more *semantic frames* that are manifested by a text, the less readable the text is. Semantic roles represent the underlying relationship of a participant with the main verb in a system. That is why the number of semantic roles presents

difficulties for readers. Named entities are also useful features for readability classification, as shown by [9].

We have derived 9 semantic indicators of text readability which are listed in Table 5. Note that the *Semafor* semantic parser [7] is used in order to annotate semantic information in Wikipedia articles. The parser uses FrameNet [12] based annotations. The Stanford named entity recognizer [13] is used for named entity parsing.

4.1.5 Other Features

The term *Hapax Legomena* is widely used in linguistics referring to words which occur only once within a context or document. These are mostly content words. Kornai [29] showed that 40% to 60% of the words in larger corpora are *Hapax Legomena*. Documents with more *Hapax Legomena* generally will contain more information. In terms of text readability, this will raise the difficulty level. Frequent content words in a corpus are considered to be familiar words. A text with more familiar words is easier to read. We extracted a list of frequent words (that occur more than 100 times) from Wikipedia. The Simple English Wikipedia[4] uses simple vocabulary and simple syntactic constructions. The vocabulary overlaps between a Wikipedia article and the simple Wikipedia will show the simpleness of an article. Our hypothesis is that a simple article will be more readable than a complex article. The probability of an article derived from an unigram model based on the Simple Wikipedia indicates the simplicity of this article. This sort of probability is calculated in a fashion similar to the approach presented in [34]. An article with a higher probability will be more readable than an article with lower probability. In this class of features, we considered 6 features: the *avg. number of hapax legomena per sentence*, the *number of hapax legomena in a text*, the *avg. number of familiar words per sentence*, the *number of familiar words in a text*, the *avg. Simple Wikipedia-related probability of a sentence* and the *Simple Wikipedia-related probability of a text*.

[4] Simple Wikipedia:http://simple.wikipedia.org/

4.2 Information-theoretic Features

Information theory measures the statistical significance of how documents vary with different types of probability distributions. What regards natural language texts, it can be used to calculate how much information can be encoded from a document based on a given probability say of words. Information theory has been developed as a statistical theory of information transmission in noisy channels. It allows us to use conditional probabilities.

4.2.1 Entropy Based Features

The most efficient way to send information through a noisy channel is at a constant rate [14, 15]. [36] have shown that this principle also correlates with biological evidence of how human language processing evolved. This rule must be retained in any kind of communication to make it efficient. In this paper we assume that any text as a medium of communication satisfies this principle. This is motivated by [14, 15] who show that some entropy rates are constant in texts. That is, for example, each sentence of a text conveys roughly the same amount of information in terms of lexical choices. The amount of information can vary in terms of the difficulty level of a text, since a more readable text differs from a less readable text in many ways. In order to utilize this information-theoretic notion we start from random variables and consider their entropy as indicators of readability.

[40] introduced entropy as a measure of information. The entropy of a random variable X is defined as

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \qquad (1)$$

The more the outcomes of X tend to a uniform distribution, the higher $H(X)$. Our hypothesis is that the higher the entropy, the less readable the text along the feature represented by X. In our experiment, we consider the following random variables: *word probability*, *character probability*, *POS probability*, *word length probability*, *lemma length probability*, *word frequency probability* (or frequency spectrum, respectively), *lemma frequency probability* and *POS frequency probability*. Note that there is a correlation between the probability distribution of words and the corresponding distribution of word

frequencies. As we use SVMs for classification, these correlations are taken into consideration.

4.2.2 Information Transmission-based Features

There is a relation among text difficulty, sentence length and word length. The usefulness of similar lexical features such as *sentence length* or *number of difficult words in a sentence* is shown in section 4.1.1. Generally a longer sentence contains more entities that influence the difficulty level. Similar things happen with longer words. But, a sentence becomes more difficult if it is longer and contains more long words. These kinds of properties can be defined by *joint* and *conditional* probabilities.

In the field of information theory, joint probability measures the likelihood of two events occurring together. That is, two random variables X and Y will be defined in the probability space. The conditional probability gives the probability that the event will occur given the knowledge that another event has already occurred. By considering the joint probability and two random variables X and Y, Shanon's joint entropy can be defined as:

$$H(X,Y) = -\sum_{<x,y>\in XxY} p(x_i,y_i)\log p(x_i,y_i) \quad (2)$$

Two conditional entropies can be defined as:

$$H(X|Y) = -\sum_{y\in Y} P(y_i)\sum_{x\in X} p(x_i|y_i)\log p(x_i|y_i) \quad (3)$$

$$H(Y|X) = -\sum_{x\in X} P(x_i)\sum_{y\in Y} p(y_i|x_i)\log p(y_i|x_i) \quad (4)$$

From the equation 1, 2, 3 and 4, it can be shown that:

$$T_s(X,Y) = H(X) + H(Y) - H(X,Y) \quad (5)$$

The function is called *Information transmission*, and it measures the strength of the relationship between elements of random variables X and Y. Details about this notion can be found in [28]. [4] used this feature to measure the amount of information about stimulus carried in a neural response. Additionally they have shown how to use this feature to validate simple stimulus-response models of neural coding of dynamic stimuli. In this work, two information transmission based features are used. These are listed in Table 8. The *Sentence length and word length probability* shows the relation between sentence length and word length and *Sentence length and complex word probability* shows the relation between sentence length and the number of complex words. The definition of a complex word is the same as a difficult word, as described in section 4.1.1. Our hypothesis is that a longer sentence with, on average, longer words or many complex words will be more difficult.

5 Experiment and Evaluation

In order to experiment with different features, the experimental data (see 1) was divided into test and training data. It should be noted that five sets of data were randomly generated where 80% of the corpus is used for training and the remaining 20% is used for testing. The weighted average of *Accuracy* and *F-score* are computed by considering all sets of data. Note that we have used the SMO [35, 25] classifier model in WEKA [19] together with the Pearson VII function-based universal kernel PUK [42].

5.1 Experiment with Linguistic Features

Since the start of readability research, lexical features have been used in many readability classification systems. Most of the traditional readability formulas (i.e., [5, 6, 26, 18, 39]) use these features. Our experimental results show that lexical features have good predictive capabilities. Table 2 shows the evaluation. The *Average vowels per document* feature represents the number of syllables in a document, and performs best among all lexical features.

The evaluation of POS-based features is presented in Table 3. The table shows that the number of *Adverbs*, *Pronouns* and *Determiners* perform better among seven POS classes investigated in this work. It has to be noted that *Determiners* are important in readability classification. That is, a determiner makes an entity more demonstrative, which might pose readability difficulties. Overall, POS-based features perform better than other linguistic features. [10] showed that content words (e.g., nouns, verbs adjective, adverbs) have higher predictive power

Table 2. Evaluation of lexical features

Features	Accuracy	F-Score
Average sentence length	37%	28%
Type-Token ratio per sentence	51%	51%
Type-Token ratio per document	55%	56%
Average difficult words per sentence	40%	31%
Number of difficult words per document	59%	59%
Type-Token ratio (lemma level) per sentence	49%	48%
Type-Token ratio (lemma level) per document	55%	55%
Average vowels per words	45%	35%
Average vowels per sentence	37%	29%
Number of vowels per document	64%	64%
10 Lexical features	71%	72%

Table 3. Evaluation of POS-based features

Features	Accuracy	F-Score
Average nouns per sentence	36%	25%
Number of nouns per document	60%	59%
Average verbs per sentence	35%	25%
Number of verbs per document	58%	55%
Average adjectives per sentence	33%	22%
Number of adjectives per document	51%	47%
Average adverb per sentence	32%	22%
Number of adverb per document	63%	63%
Average pronouns per sentence	38%	29%
Number of pronouns per document	63%	63%
Average preposition per sentence	38%	30%
Number of preposition per document	54%	49%
Average determiners per sentence	36%	27%
Number of determiners per document	63%	63%
15 POS-based features	71%	71%

than function words (e.g., pronouns, preposition, grammatical article). However, the experimental results show that function words like pronoun and definite article should also be considered when measuring the difficulty of a text.

There is a correlation between readability difficulties and syntactical complexity. Table 4 shows the evaluation of syntactical features. The *average number of noun phrases per document* is a good predictive feature, which confirms [3]. [34] showed that a sentence with subordinate clauses is difficult for a less-skilled reader. However, our experimental results show that other syntactical features are more important than subordinate clauses.

A text is said to be easier for readers when the underlying semantics of the text are easier to understand. So readability difficulty is influenced by the semantic entities of an article. Table 5 shows the evaluation of semantic based features. [9] noted that a named entity is a cognitively motivated feature, and an article with many named entities is difficult for people with intellectual disabilities. But that is not reflected by our findings. The *Average co-reference chain length* represents the average

Table 4. Evaluation of syntax-based features

Features	Accuracy	F-Score
Average noun phrases per sent.	36%	26%
Number of phrases per document	62%	62%
Average verb phrases per sent.	34%	25%
Number of verb phrases per document	59%	55%
Average prepositional phrases per sent.	37%	28%
Number of prepositional phrases per document	61%	61%
Average length of subordinate clauses per sent.	34%	24%
Number of subordinate clauses per document	52%	45%
Average embedded clauses per sent.	35%	27%
Number of embedded clauses per document	59%	56%
10 Syntax features	67%	67%

Table 5. Evaluation of semantic-based features

Features	Accuracy	F-Score
Average named entities per sent.	35%	32%
Number of named entities per document	48%	45%
Number of co-reference chains in a document	62%	63%
Average co-reference chain length	43%	40%
Average distance antecedent-anaphora	54%	51%
Average number of semantic frames per sent.	36%	27%
Number of semantic frames per document	63%	63%
Average number of semantic roles per sent.	34%	24%
Number of semantic roles per document	65%	65%
9 semantic features	69%	68%

number of noun phrases that refer to the same entity. Our hypothesis was that a document with a longer co-reference chain would be more difficult. Readers have to keep these entities in memory in order to understand the relation between them. However, the results in Table 5 do not support this hypothesis. Semantic frames and semantic roles have good predictive capabilities for readability classification.

Table 6 shows the evaluation of some of the orthodox features. The *number of familiar words per document* is one of the best performing individual features. That is, an article with more known (frequent) words is more readable. Vocabulary overlap between a Wikipedia article and the *Simple Wikipedia* unigram model is also a good indicator of readability. Simple Wikipedia can be used to build a language model for a similar kind of task. Table 6 shows the combined result of all linguistic features, and demonstrates that the 49 linguistic features we used give better results than those reported by many previous research projects (as noted in section 2).

Table 6. Evaluation of other features

Features	Accuracy	F-Score
Average Hapax-legomena per sentence	51%	50%
Number of Hapax-legomena per document	57%	53%
Average number of familiar words per sentence	34%	23%
Number of familiar words per document	65%	65%
Average number of simple words per sentence	37%	28%
Number of simple words per document	63%	63%
6 other features	68%	68%
49 linguistic features	73%	73%

Table 7. Evaluation of entropy based features

Features	Accuracy	F-Score
Word probability	55%	50%
Character probability	40%	37%
POS probability	35%	25%
Word length probability	42%	33%
Word frequency probability	50%	49%
Lemma length probability	40%	32%
Lemma frequency probability	51%	50%
Character frequency probability	47%	43%
POS frequency probability	51%	45%
9 Entropy features	72%	72%

5.2 Experiment with Information-theoretic Features

As noted earlier, entropy measures the amount of information in an article. Wikipedia's articles are assumed to be a medium of communication between Wikipedians and readers. Conversely, information flow of a readable article will differ from that of a less readable article. Thus, the constants for the corresponding entropy rates of the different readability classes will differ. As a single feature, these entropy-based features perform as well as linguistic features. But when considered collectively, these are the best performing feature set among all feature sets. Among all similar features the random variable with *Word Probability* works better than others. It should be noted that some of the entropy based features are linguistically motivated. That is, probabilities are calculated from the output of some of the linguistic tools. Table 7 shows the evaluation.

Table 8 shows the evaluation of information transmission based features. These features show the relationship of word length and number of complex words with the sentence length. Individually these features perform better than many other individual features. In total 11 information-theoretic features are used, which perform similarly to 49 linguistically motivated features. We get 75% *accuracy* and 75% *F-score* when these 11 information-theoretic features are added to the 49 linguistic features.

Table 8. Evaluation of information-transmission based features

Features	Accuracy	F-Score
Sentence length and word length probability	64%	64%
Sentence length and complex word probability	60%	60%
Information transmission based features	63%	63%
11 Information theoretic features	73%	73%
60 linguistic + information-theoretic features	75%	75%

6 Discussion

Table 1 shows that an article of a higher readability class is often longer than an article with lower readability class. However, this feature is not reflected in our experiment. The *average sentence length* is one of the worst performing features. The POS-based features outperforms syntax based features, which supports the findings in [10]. It should also be noted that lexical features are better predictors than many linguistic features. That is why traditional readability formulas are still considered in many commercial readability assessment tools. Semantic features are more predictive of readability classification. But note that the number of named entities feature did not perform as we expected. The result in this work suggests that readability difficulty is influenced by the underlying semantics of an article. The *number of semantic roles in a document* and *the number of familiar words* are the best performing individual features. The evaluation of all linguistic features suggest that features should be calculated per document instead of per sentence for a document-level readability classification.

The result of entropy-based features showed that a written document also might have a constant entropy rate, as shown in [14, 15]. We found that 11 information-theoretic features performed as well as 49 linguistically motivated features. Many languages are considered to be low-density languages, either because the population speaking the language is not very large, or because insufficient digitized text material is available in the language even though millions of people speak the language. These information-theoretic concepts should be considered in order to build a readability classifier for these languages.

7 Conclusion and Future Work

We compared 49 linguistic indicators of readability with 11 information-theoretic features in order to separate the text into different classes of text readability. The latter features represent the amount of information of a text along some of its simple quantitative characteristics. Our experimental results show that information-theoretic features perform as well as their linguistic counterparts. They provide an easy way to compute readability, in contrast to the much more complex linguistic variables considered here. Note that POS-based linguistic features outperform syntactic and semantic features, but lexical features still dominate. From this perspective one may say that readability classification does not require complex measurements. Note also that the combination of linguistic and information-theoretic features outperforms these feature sets when considered in isolation. Thus, the two feature sets seem to measure slightly different things. To the best of our knowledge, the information-theoretic features considered here were not explored in previous studies of text readability classification. In future work, we will extend the information-theoretic approach by using larger corpora of crowd-sourced readability data. We expect to get this data step by step due to the collaborative principle by which Wikipedia grows.

Acknowledgements

We would like to thank Andy Lücking, Armin Hoenen and Paul Warner for their fruitful suggestions and comments. We also thank three anonymous reviewers. This work is funded by the LOEWE Digital-Humanities project in the Goethe-Universität Frankfurt.

References

1. **Alan, K.** (**2001**). *Natural Language Semantics*. Blackwell Publishers Ltd, Oxford.
2. **Aluisio, R., Specia, L., Gasperin, C., & Scarton, C.** (**2010**). Readability assessment for text simplification. In *NAACL-HLT 2010: The 5th Workshop on Innovative Use of NLP for Building Educational Applications*.
3. **Barzilay, R. & Lapata, M.** (**2008**). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 21(3), 285–301.
4. **Borst, A. & Theunissen, F. E.** (**1999**). Information theory and neural coding. *Nature Neuroscience*, 2, 947–957.
5. **Dale, E. & Chall, J. S.** (**1948**). A formula for predicting readability. *Educational Research Bulletin*, 27(1), 11–20+28.
6. **Dale, E. & Chall, J. S.** (**1995**). *Readability Revisited: The New Dale-Chall Readability formula*. Brookline Books.
7. **Das, D. & Smith, N. A.** (**2011**). Semi-supervised frame-semantic parsing for unknown predicates. In *The Annual Meeting of the Association for Computational Linguistics, Portland*.
8. **Eickhoff, C., Serdyukov, P., & de Vries, A. P.** (**2011**). A combined topical/non-topical approach to identifying web sites for children. In *Proceedings of the fourth ACM international conference on Web search and data mining*.
9. **Feng, L., Elhadad, N., & Huenerfauth, M.** (**2009**). Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL*.
10. **Feng, L., Janche, M., Huenerfauth, M., & Elhadad, N.** (**2010**). A comparison of features for automatic readability assessment. In *The 23rd International Conference on Computational Linguistics (COLING)*.
11. **Fillmore, C. J.** (**1982**). Frame semantis. In *Linguistics in the Morning Calm*. Hanshin Publishing Co., 111–137.
12. **Fillmore, C. J., Johnson, C. R., & Petruck, M. R.** (**2003**). Background to framenet. *International Journal of Lexicography*, 16(3), 235–250.
13. **Finkel, J. R., Grenager, T., & Manning, C.** (**2005**). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*.
14. **Genzel, D. & Charniak, E.** (**2002**). Entropy rate constancy in text. In *Proceedings of the 40st Meeting of the Association for Computational Linguistics (ACL 2002)*.
15. **Genzel, D. & Charniak, E.** (**2003**). Variation of entropy and parse trees of sentences as a function of the sentence number. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

16. **Gildea, D. & Jurafsky, D.** (**2002**). Automatic labeling of semantic roles. *Computational Linguistics*, 28(3), 245–288.
17. **Giles, J.** (**2005**). Internet encyclopaedias go head to head. *Nature*, 438, 900:901.
18. **Gunning, R.** (**1952**). *The Technique of clear writing*. McGraw-Hill; Fourh Printing Edition.
19. **Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H.** (**2009**). The WEKA data mining software: an update. *ACM SIGKDD Explorations*, 11(1), 10–18.
20. **Heilman, M., Collins-Thompson, K., & Eskenazi, M.** (**2007**). Combining lexical and grammatical features to improve readavility measures for first and second language text. In *Proceedings of the Human Language Technology Conference*.
21. **Heilman, M., Collins-Thompson, K., & Eskenazi, M.** (**2008**). An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications (EANL)*.
22. **Holler, A. & Irmen, L.** (**2007**). Empirically assessing the effects of the Right Frontier Constraint. In **Branco, A.**, editor, *Anaphora: Analysis, Algorithms and Applications*, Lecture Notes in Artificial Intelligence. Springer, Berlin and Heidelberg, 15–27.
23. **Islam, Z., Mehler, A., & Rahman, R.** (**2012**). Text readability classification of textbooks of a low-resource language. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation. (Accepted)*.
24. **Kate, R. J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R. J., Roukos, S., & Welty, C.** (**2010**). Learning to predict readability using diverse linguistic features. In *23rd International Conference on Computational Linguistics (COLING 2010)*.
25. **Keerthi, S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K.** (**2001**). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, 13(3), 637–649.
26. **Kincaid, J., Fishburne, R., Rodegers, R., & Chissom, B.** (**1975**). Derivation of new readability formulas for Navy enlisted personnel. Technical report, US Navy, Branch Report 8-75, Cheif of Naval Traning, Millington, TN.
27. **Klein, D. & Manning, C. D.** (**2003**). Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL 2003)*.
28. **Klir, G. J.** (**2005**). *Uncertainty and Information*. Wiley-Interscience.
29. **Kornai, A.** (**2008**). *Mathematical Linguistics*. Springer.
30. **Ma, Y., Singh, R., Fosler-Lussier, E., & Lofthus, R.** (**2012**). Comparing human versus automatic feature extraction for fine-grained elementary readability assesment. In *NAACL-HLT 2012 Workshop on Predicting and Improving Text Readability for target reader populations*.
31. **Mann, W. & Thompson, S.** (**1988**). Rhethorical structure theory: Towards a functional theory of text organization. *Text*, 8(3), 243–281.
32. **Mullan, W.** (**2008**). Dairy science and food technology improving your writing using a readability calculator.
33. **Petersen, S. E. & Ostendorf, M.** (**2009**). A machine learning approach to reading level assesment. *Computer Speech and Language*, 23(1), 89–106.
34. **Pitler, E. & Nenkova, A.** (**2008**). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
35. **Platt, J. C.** (**1998**). *Fast training of support vector machines using sequential minimal optimization*. MIT Press.
36. **Plotkin, J. B. & Nowak, M. A.** (**2000**). Language evolution and information theory. *Journal of Theoretical Biology*, 205(1), 147–159.
37. **Polanyi, L.** (**1988**). A formal model of the structure of discourse. *Journal of Pragmatics*, 12(56), 601–638.
38. **Schwarm, S. E. & Ostendorf, M.** (**2005**). Reading level assessment using support vector machines and statistical language models. In *the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics(ACL 2005)*.
39. **Senter, R. & Smith, E. A.** (**1967**). Automated readability index. Technical report, Wright-Patterson Air Force Base.
40. **Shannon, C. E.** (**1948**). A mathematical theory of communication. *The Bell System Technical Journal*, 27(1), 379–423.
41. **Stoyanov, V., Cardie, C., Gilbert, N., Riloff, E., Buttler, D., & Hysom, D.** (**2010**). Coreference resolution with reconcile. In *Conference of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Short Paper*.
42. **Üstün, B., Melssen, W., & Buydens, L.** (**2006**). Facilitating the application of support vector regression by using a universal Pearson VII function based kernel. *Chemometrics and Intelligent Laboratory Systems*, 81(1), 29–40.

Alexander Mehler is professor for Text Technology at the Goethe University Frankfurt am Main, where he heads the Text Technology Lab as part of the Institute of Computer Science. Alexander Mehler is also member of the executive committee of the LOEWE Priority Program *Digital Humanities* at the Goethe University. His research interests include the empirical analysis and simulative synthesis of discourse units in spoken and written communication. He aims at a quantitative theory of networking in linguistic systems to enable multi-agent simulations of their life cycle. Alexander Mehler integrates models of semantic spaces with simulation models of language evolution and topological models of network theory to capture the complexity of linguistic information systems. Currently, he is heading several research projects on the analysis of linguistic networks.

Zahurul Islam has been working in the Text Technology Lab at the Goethe University, Frankfurt as a research assistance since February 2011. He is also a PhD candidate at the Institute of Computer Science, Goethe University, Frankfurt. Before joining the Text Technology Lab, he worked as a graduate researcher at the Department of Computer Science, University of Pisa. Zahurul Islam studied language technology at the University of Saarland, Germany; linguistics at the Groningen University, The Netherlands; and computer science at the BRAC University, Bangladesh. His research interests include text readability classification, statistical natural language processing, machine learning and machine translation.

Article received on 12/12/2012; accepted on 16/02/2013.