Readability-based Sentence Ranking for Evaluating Text Simplification

Sowmya Vajjala
Iowa State University, USA
sowmya@iastate.edu

Detmar Meurers
University of Tuebingen, Germany
dm@sfs.uni-tuebingen.de

Abstract

We propose a new method for evaluating the readability of simplified sentences through pair-wise ranking. The validity of the method is established through in-corpus and cross-corpus evaluation experiments. The approach correctly identifies the ranking of simplified and unsimplified sentences in terms of their reading level with an accuracy of over 80%, significantly outperforming previous results.

To gain qualitative insights into the nature of simplification at the sentence level, we studied the impact of specific linguistic features. We empirically confirm that both word-level and syntactic features play a role in comparing the degree of simplification of authentic data.

To carry out this research, we created a new sentence-aligned corpus from professionally simplified news articles. The new corpus resource enriches the empirical basis of sentence-level simplification research, which so far relied on a single resource. Most importantly, it facilitates cross-corpus evaluation for simplification, a key step towards generalizable results.

1 Introduction

Text simplification is the process of simplifying the linguistic form of a text without losing its meaning. It has applications in several domains ranging from language learning (Petersen and Ostendorf, 2007) and biomedical information extraction (Jonnalagadda and Gonzalez, 2010) for human readers all the way to automatic simplification designed to improve parsing by machines (Chandrasekar et al., 1996). While manual simplification relies entirely on expert writers, semi-automatic approaches serve as an assistive tool for writers, alerting them of text passages that may be difficult to read for the target audience and indicating how to rewrite them (Candido et al., 2009). Automatic text simplification approaches, generating simplified text from an unsimplified version by means of hand-crafted rules, data-driven methods, or hybrid techniques have also been proposed (e.g., Woodsend and Lapata, 2011; Siddharthan and Mandya, 2014).

The nature of the simplification performed depends on the purpose of the approach. Thus, the evaluation of a system that aims to improve the parser speed (Chandrasekar et al., 1996) also differs from one that was developed to support spoken language understanding (Tur et al., 2011)In an educational context, typically the purpose is to adapt the text to a level of complexity facilitating comprehension by the target audience, such as language learners or students at a particular grade level. It thus is important to be able to evaluate the complexity of simplified and unsimplified versions of a text – which is the issue we address in this paper. The approach is equally applicable to identifying those parts of a text that are particularly complex and thus constitute good targets for simplification.

Text simplification is generally evaluated in one of three ways: through small-scale user evaluations, with a machine translation metric, or using readability measures (Siddharthan, 2014). We explore the third option. Evaluating text simplification using readability assessment is typically carried out with traditional readability formulae. For example, Woodsend and Lapata (2011) make use of the Flesch-Kincaid Reading Ease formula. Such readability formulae are based on surface-level features: the average sentence length, word length, or lists of difficult words (cf. DuBay, 2006). While such features often correlate with the actual causes of difficulty in a piece of text (e.g., complex syntax, infrequent words), manipulating these surface features does not necessarily result in more readable texts (cf. Klare, 1974); they merely are surface indicators of a broad range of underlying linguistic and psychological characteristics of authentic texts targeting different audiences. Recent research also showed that more sophisticated, linguistically-grounded models support a more reliable assessment of readability (e.g., Nelson et al., 2012). Although readability assessment is primarily studied at a text-level, recent research explored it for sentences as well (e.g., Napoles and Dredze, 2010; Medero and Ostendorf, 2011; Pilán et al., 2014; Dell'Orletta et al., 2014; Vajjala and Meurers, 2014a). Being able to assess the readability at the sentence level is important to identify targets for simplification and to quantify the degree of simplification performed for a given sentence.

Vajjala and Meurers (2014a) studied sentential readability for text simplification. They show that a relative comparison instead of an absolute classification is better suited to identifying the difficult sentences compared to simplified versions. They used a regression model trained on whole documents to compare the readability of parallel sentences from the sentence-aligned Wiki-SimpleWiki corpus (Zhu et al., 2010). While agreeing with the general perspective, we propose a ranking approach which more directly captures the idea of relative levels of readability, and we show that it significantly outperforms the state of the art for evaluating sentential simplification.

To support an evaluation of sentential simplification across different corpus resources, testing whether something general has been learned, we created a new corpus of aligned simple-complex sentence pairs. We collected the data from the web site OneStopEnglish.com, which offers manually simplified versions of news articles for language learners. We tested our approach with this corpus, with the standard Wikipedia-SimpleWikipedia sentence-aligned corpus, and using cross-corpus evaluation. We show that our approach outperforms previous approaches for identifying sentential simplifications and that the performance generalizes across corpora.

In terms of a qualitative analysis, we compare groups of features in terms of their contribution to the ranking model and find that both word-level and sentence-level properties play a role in ranking the sentences by their reading level. While the psycholinguistic measures of word properties figure prominently among the top features, the best-performing model consists of all the features.

In sum, the contributions of this paper are:

1. We propose an approach to evaluate text simplification methods in terms of reading levels. Through multiple cross-corpus tests, we show that the approach performs at an accuracy of over 80%.

2. We compiled a new corpus of sentence-level simplifications to obtain a broader empirical basis on which to evaluate and train text simplification systems.

3. We explored the role of individual features and feature groups for this task, including a comparison them across corpora.

4. In terms of the practical application context we are targeting, the quantitative and qualitative results in this paper establish that the approach can meaningfully be used in practice to evaluate simplification systems developed with the aim of reducing the difficulty of informative text for language learners.

The paper is organized as follows: Sections 2 and 3 describe the corpora and feature set we used. Section 4 discusses our experiments and their results. Sections 5 and 6 put our research in context and conclude the paper.

2 Corpora

The practical purpose of our approach is to evaluate text simplification approaches aimed at helping language learners. Hence we train and test our models on two corpora created with this target audience in mind.

2.1 OneStopEnglish corpus

OneStopEnglish (OSE) is a resource website for English teachers published by the Macmillan Education Group. They offer Weekly News Lessons[1] consisting of news articles sourced from the newspaper *The Guardian*. The articles are rewritten by teaching experts in a way targeting English language learners at three reading levels (elementary, intermediate, advanced). We obtained permission to use the articles for research purposes and downloaded the weekly lessons from September 2012–March 2014, which resulted in a collection of 76

[1] `http://www.onestopenglish.com/skills/news-lessons/weekly-topical-news-lessons`

article triplets. Each article is included in its elementary, its intermediate, and its advanced version so that overall the corpus contains 228 articles.

Corpus pre-processing The weekly lessons are pdf files consisting of a pre-test about the topic of the article, the re-written news article, and exercises related to the article. We first parsed the pdfs using iTextPDF[2] to extract the article text, excluding everything else. Since our aim is to compare different versions of a sentence, we took each article triplet and sentence-aligned two at a time using TF-IDF and cosine similarity, following previous research on monolingual sentence alignment (Nelken and Shieber, 2006; Zhu et al., 2010).

OSE3 For the sentences which exist in all three versions of an article (elementary, intermediate, advanced), we obtain a triplet of sentences. We selected all triplets for which each pair of sentences differed and was above a minimum similarity threshold of 0.7 (based on manual qualitative analysis using different thresholds). Overall, we identified 837 sentence triplets and refer to this corpus as OSE3.[3] An example of a sentence rewritten across the three levels is shown below:

Adv: *In Beijing, mourners and admirers made their way to lay flowers and light candles at the Apple Store.*

Int: *In Beijing, mourners and admirers came to lay flowers and light candles at the Apple Store.*

Ele: *In Beijing, people went to the Apple Store with flowers and candles.*

OSE2 We compiled a second corpus consisting of pairs of sentences, which we will refer to as OSE2. We extracted the 3113 pairs of sentences (elementary-intermediate, intermediate-advanced, or elementary-advanced) that differed and were above the minimum similarity threshold.

2.2 Wikipedia-SimpleWiki corpus

Simple English Wikipedia (SimpleWiki) targets children and adults who are learning English,[4] so a corpus of sentence pairs from Wikipedia and SimpleWiki suits our goal to compare sentence-level text simplification. We use the sentence-aligned corpus created by Zhu et al. (2010). They compiled a collection of ∼65k parallel articles from Wikipedia and SimpleWiki to create a sentence-aligned corpus consisting of ∼100k pairs. We used a subset of this corpus consisting of 80,912 sentence pairs, after removing the sentence pairs that are identical in both versions.

3 Features and Methods

3.1 Features

Vajjala and Meurers (2014a) introduce a range of lexical, syntactic, morphological, and psycholinguistic features to build a document-level readability model that performed on par with existing commercial and academic systems on the Common Core State Standard test set for English (CCSSO, 2010). They show that the model can also be applied at the sentence-level.

Given our goal of studying the effectiveness of ranking over regression and the relevance of specific features for sentence-level ranking, we built our models using the same feature set they used, which makes a direct comparison possible. The feature set of consists of four groups of features:

- LEX consists of lexical diversity and density features primarily based on type-token and POS ratios, inspired by Second Language Acquisition (SLA) research, and lexical semantic properties from WordNet (Miller, 1995) such as the average number of senses of a word.
- SYN includes syntactic features based on specific patterns extracted from parse trees, including measures of syntactic complexity from SLA research.
- CEL is a group of features encoding morpho-syntactic properties of lemmas, estimated using the Celex (Baayen et al., 1995) database.
- PSY contains word-level psycholinguistic features such as concreteness, meaningfulness and imageability extracted from the MRC psycholinguistic database (Wilson, 1988) and various Age of Acquisition (AoA) measures released by Kuperman et al. (2012).

3.2 Methods

We model sentential complexity as a pair-wise ranking problem. Pair-wise ranking is one of the

[2] `http://itextpdf.com`

[3] We will share our sentence-aligned corpora for research purposes under a standard CC BY NC SA license.

[4] `http://simple.wikipedia.org/wiki/Simple_English_Wikipedia`

learning-to-rank approaches, typically used in information retrieval for ranking search results (Li, 2014). In that scenario, it is used to compare a pair of documents in terms of their relevance to a given query. In our case, the aim of the ranker given a pair of sentences (where one is the simplified version of the other) is to predict which one of them is simpler than the other. Thus, the learning problem for us is to compare sentence pairs within a group of sentences and rank them based on their complexity, trying to minimize inversion of ranks.

To apply ranking, we need to have a numeric score for (the feature vector of) each sentence. In Wiki and OSE2, we assigned a reading level of 2 to the more difficult version and 1 to its simplified version in the sentence pair. For the sentence triplets in OSE3, we used the sentences scores 3 (advanced), 2 (intermediate), and 1 (elementary). Since pair-wise ranking only considers relative ranks, the ranking procedure is not dependent on the specific absolute reading level of a sentence. In the case of sentences that were split into two in the simplified version, we scored both the simple sentences as 1 so that no pair-wise constraints are generated between them.

Ranking: We explored three ranking algorithms.

RankSVM (Herbrich et al., 2000) transforms ranking into a pair-wise classification problem and uses a Support Vector Machine for learning to rank the sentence pairs. It is one of the most commonly used ranking algorithms in NLP tasks. and was also employed in a related task, for ranking children's literature texts based on their reading level (Ma et al., 2012).

RankNet (Burges et al., 2005) is a pair-wise ranking algorithm that is a modified version of a traditional back-propagation-based neural network, applied to ranking problems. It is known to perform well in practice and was successfully used in a real-life search engine to rank search results.

RankBoost (Freund et al., 2003) is an algorithm that uses boosting for pair-wise ranking. It uses a linear combination of several weak rankers to produce the final ranking. The algorithm was typically used in collaborative filtering problems.

We used publicly available implementations of these algorithms for training our models: SVM^{rank} (Joachims, 2006)[5] for RankSVM and RankLib[6] software for RankNet and RankBoost.

Evaluation: Since our learning goal is to minimize the number of wrongly ranked pairs, we evaluate the approach in terms of the percentage of correctly ordered pairs. In other words, we report the percentage of pairs in which the difficult version gets a higher rank than its simplified counterpart. We refer to this as accuracy.

4 Experiments

4.1 Reference performance on Wiki dataset

We start with an experiment directly comparing the ranking approach with the results reported in (Vajjala and Meurers, 2014a) on the Wiki-SimpleWiki data set. They used a regression model trained on documents to get the reading levels for sentences. Their model identified the rank order correctly in 59% of the cases and assigned equal score to both versions of the sentence in 11% of the cases. So, randomly considering half of the 11% cases as correctly ordered and half as wrongly ordered, one obtains a ranking accuracy of 64.5% for their model.

Replacing regression with ranking, we trained a model using SVM^{Rank} on the entire Wiki dataset in a 10-fold cross validation (CV) setup. The ranking model achieves an accuracy of 82.7%, which is a significant improvement over the baseline (p < 0.01). The Standard Deviation between the ten folds is 8.4%. This high level of variability suggests that the nature of what constitutes simplifications in SimpleWiki varies significantly, as may be expected for a collaborative editing setup – a potentially interesting issue to explore in the future.

As several text simplification approaches (Zhu et al., 2010; Woodsend and Lapata, 2011) used the Flesch-Kincaid Grade Level formula, which is based on the average word and sentence length as surface features, as a readability measure for text simplification, we also determined the accuracy of ranking the sentences in the Wiki-SimpleWiki data using this formula and obtained an accuracy of 72.3%.

As summed up in Table 1, on the Wiki-SimpleWiki dataset the ranking approach clearly

[5] `http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html`

[6] `http://sourceforge.net/p/lemur/wiki/RankLib`

outperforms the regression approach of Vajjala and Meurers (2014a) and the Flesch-Kincaid readability formula. The rich linguistic feature set we have adapted from Vajjala and Meurers (2014a) thus can clearly outperform the surface-based readability formula, but the richer information only becomes effective when the relative readability of pairs of sentences is learned using a dedicated ranking algorithm.

Approach	Accuracy
Vajjala and Meurers (2014a)	64.5%
Flesch-Kincaid formula	72.3%
Our RankSVM approach	82.7%

Table 1: Ranking accuracy on Wiki-SimpleWiki

To explore things further, we next compared different ranking approaches and tested the generalizability of the ranking approach in a cross-corpus setup and in multi-level simplification scenarios.

4.2 Ranking algorithms and generalizability

To compare the three ranking algorithms introduced in Section 3.2, we first trained ranking models for the WIKI and OSE2 corpora. To make the results comparable for these two corpora, we selected 2,000 sentence pairs for each of the training sets (WIKI-TRAIN, OSE2-TRAIN), and used the remaining part as the test set (WIKI-TEST: 78,912 pairs, OSE2-TEST: 1,113 pairs).

Table 2 presents the performance of the three ranking algorithms on the two training sets for within and cross-corpus evaluation. As a baseline, the Flesch-Kincaid formula results in 69.0% for WIKI-TEST and 69.6% for OSE2-TEST.

TRAIN	TEST	RankSVM	RankNet	RankBoost
WIKI	WIKI	**81.8%**	72.5%	76.4%
WIKI	OSE2	74.6%	59.1%	70.2%
OSE2	WIKI	77.5%	73.8%	74.8%
OSE2	OSE2	**81.5%**	69%	75.5%

Table 2: Accuracies for the three rank algorithms

RankSVM performs best among the ranking algorithms we tried. This also generalizes to the cross-corpus tests. In the following experiments, we therefore only report the results for RankSVM.

Cross-corpus evaluation always shows a drop in performance. The drop is smaller for the model trained on the OSE2 corpus, which suggests that the OSE2 corpus covers a broader, more representative range of simplifications. Taking that idea further, we explored improving cross-corpus performance using two methods enriching the training data.

4.3 Improving cross-corpus performance

First, we combined the two training sets to create a hybrid training set WIKI-OSE2-TRAIN, which should increase the representativity and range of the simplifications included in the training data.

Second, we used the three level corpus OSE3 to train the ranker to simultaneously consider a broader range of simplifications: the ranker will learn a single set of weights for ranking the three pairs in a set for OSE3, instead of three sets of weights for ranking each pair independently. We randomly selected 750 sentence triplets from the OSE3 corpus as training set (OSE3-TRAIN), leaving the remaining 87 as held-out test set (OSE3-TEST). Table 3 shows the results on the three test sets for the models trained on WIKI-OSE2-TRAIN and OSE3-TRAIN. The baseline accuracy obtained using the Flesch-Kincaid formula for OSE3-TEST is 71.6%.

	WIKI-OSE2-TRAIN	OSE3-TRAIN
WIKI-TEST	81.3%	78.6%
OSE2-TEST	80.7%	82.4%
OSE3-TEST	79.7%	79.7%[7]

Table 3: Accuracies for the extended training sets

As expected, the accuracy for the combined, more varied training set WIKI-OSE2-TRAIN results in a comparable performance across the three tests sets. It seems to account for a broader range of simplification options. The results for the OSE3-TRAIN set, providing the ranker with triples over which to learn the weights, are less clear.

Overall, the fact that cross-corpus and same-corpus results are relatively close together supports the assumption that reliable sentence-level readability ranking models which generalize across very different data sets can be built.

4.4 Influence of the amount of training data

While the WIKI-OSE2-TRAIN contains more diverse data, it also contains twice as much data as the two smaller training sets it consists of. To isolate the effect of the training data size, we explored how the low cross-corpus performance of 74.6% of for the WIKI-TRAIN model on the OSE2-TEST

[7] The identical overall performance of both models on the OSE3-TEST set differs in terms of individual instances.

we saw in Table 2 is improved simply by increasing the amount of training data. We therefore trained on increasingly larger portions of the Wiki-SimpleWiki data set up to the full set of 80k pairs. We tested on OSE2-TEST and additionally on OSE3-TEST to lower the risk of idiosyncrasies specific to a single test set. Figure 1 shows the accuracies for models trained on increasing amount of Wiki-SimpleWiki training data.

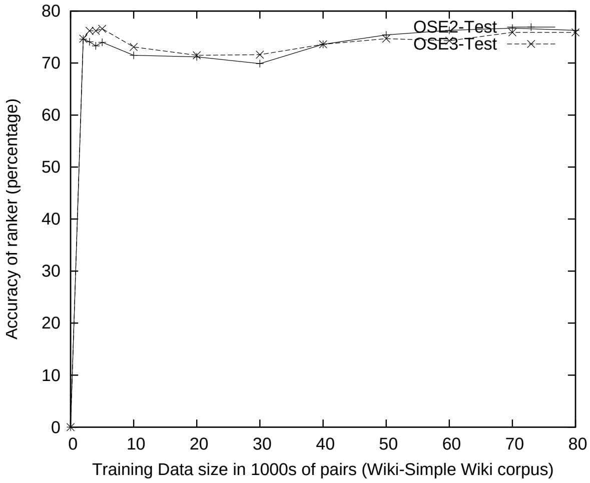

Figure 1: Accuracies for increasing training size

The curve is essentially flat, with the model on the largest training set (80k) reaching an accuracy of 76.3%, less than two percent above the result we obtained using only 2k pairs for training, and significantly below the 80.7% obtained for the model trained on the 4k WIKI-OSE2-TRAIN set (cf. Table 3). The Wiki-SimpleWiki data thus does not seem to offer the variety of simplification needed to generalize better to the OSE test sets.

4.5 Feature Selection

Turning from experiments establishing the overall validity of the approach to the impact of the different features, the next experiments explore feature selection. In addition to characterizing how much can be achieved with how little, feature selection gives us an opportunity to better understand the linguistic characteristics of simplification. We explored which features contribute the most as single features or as feature groups.

Impact of feature groups: First, we investigated the contribution of different feature groups to ranking accuracy. Figure 2 presents the performance of ranking models trained using the four feature groups (LEX, SYN, CEL, PSY) and a model trained with all features. We train on the WIKI-OSE2-TRAIN dataset, which as we saw in Table 3 generalized well across different test sets.

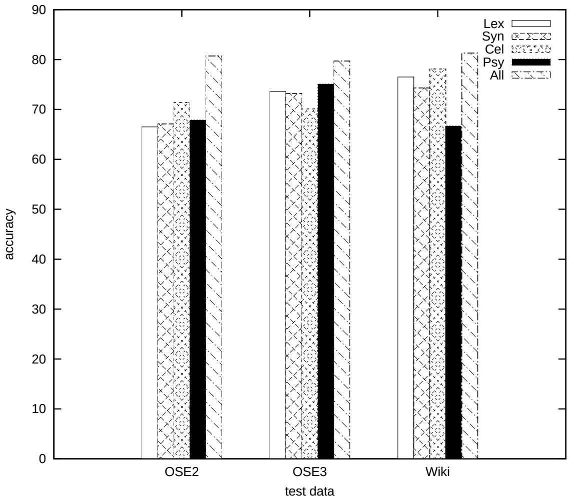

Figure 2: Performance of different feature groups

Figure 2 shows that the performance of the individual feature groups varies with the test-sets used. For example, CEL features result in lower accuracy for OSE3-TEST compared to other feature groups, whereas PSY features performed poorly for WIKI. For all test sets, the model trained with all the features outperforms the individual feature group models. For a generalizing approach to the evaluation of text simplification, modeling multiple dimensions of readability thus is more useful than choosing a single aspect.

Impact of individual features: To understand the linguistic characteristics of simplification, it is useful to identify which individual features are more informative for these authentic data sets. Hence, we trained single feature ranking models with each of the training sets and ranked the features based on their performance on the test sets. The list of single features achieving over 60% for in-corpus test-set evaluation are shown in Table 4 for the WIKI data and Table 5 for the OSE2 data.[8]

For the WIKI data only six features individually performed with an accuracy above 60%: four SYN, one LEX, one PSY. For the OSE2 data, this was the case for eight features: two SYN, one LEX, five PSY. Both lists thus include a combination of word-level and syntactic features, with syn-

[8] We experimented with a range of AoA norms and lexical diversity measures, but for space reasons include only the best AoA and lexical diversity feature here. Interestingly, the accuracies obtained for the various AoA norms differed substantially, between 37% and 72.8%, also due to coverage.

Feature	**Group**	**Accur.**
num. subtrees (SUBTREES)	SYN	72.1%
corrected type-token ratio (CTTR)	LEX	70.4%
sentence length (SENLEN)	SYN	69.7%
avg. Age of Acquisition Kup.-Lem. (AoA)	PSY	64.8%
num. constituents per tree (CONST)	SYN	63.3%
avg. length of t-unit (MLT)	SYN	63.2%

Table 4: Performance of single feature ranking models for Wiki-Train/Wiki-Test

Feature	**Group**	**Accur.**
AoA	PSY	72.8%
CTTR	LEX	66.7%
SUBTREES	SYN	64.4%
avg. length of clause (MLC)	SYN	63.2%
avg. word imagery rating (IMAGERY)	PSY	63.2%
avg. word familiarity rating (FAMILIARITY)	PSY	63.2%
avg. Colorado meaningfulness rating of words (MEANINGFULNESS)	PSY	63.2%
avg. concreteness rating (CONCRETENESS)	PSY	61.7%

Table 5: Performance of single feature ranking models for OSE2-Train/OSE2-Test

tactic simplification playing more of a role for the Wiki dataset and lexical choices relating to psycholinguistic characteristics being more relevant for the OSE2 data.

Sentence length is more predictive for Wiki than for OSE2, probably because the Wiki dataset contains a lot of deletions ($\sim$45% of the sentence pairs show major deletions) compared to the OSE dataset, where sentences were mostly rewritten or paraphrased instead of deleting content. In line with this observation, sentence length as a single feature for OSE2-Test data achieves an accuracy of only 57.5%, compared to the 69.7% for Wiki-Test.

The role and interdependence of the different psycholinguistically motivated features (age of acquisition, concreteness, meaningfulness, imagery) for the OSE2 data is interesting and would merit further study. A good understanding of the role of these features would be directly relevant for improving lexical simplification approaches such as that of Jauhar and Specia (2012), which already integrates some features from the MRC psycholinguistic database to rank word substitutes for lexical simplification.

4.6 Simplification at different levels

We next explored, whether the nature of the simplification differs between advanced sentences being simplified compared to intermediate sentences being (further) simplified. To investigate this, we split the OSE3-Train and OSE3-Test datasets into two pairs of datasets Adv-Int-Train, Adv-Int-Test and Int-Ele-Train, Intr-Ele-Test. Table 6 shows the differences in the performance of the ranking approach between the two levels of simplification.

	Training Data	
	Adv-Int	Int-Ele
Adv-Int-Test	73.6%	74.7%
Int-Ele-Test	81.6%	80.5%

Table 6: Simplification at different levels

The performance on the Int-Ele-Test set is better than that on the Adv-Ele-Test set, independent of whether the model was trained on the Adv-Int or Int-Ele training data. To understand the reason, we explored the nature of the simplification involved at these two different levels by testing the predictive power of individual features.

Table 7 shows the list of features that individually achieved an accuracy of over 60% for intermediate to beginner level simplification. For advanced to intermediate, only AoA features achieved an accuracy of above 60%.

The better overall performance at the intermediate to elementary simplification level and the higher number of informative features at that level indicate that the nature of the simplification between advanced and intermediate sentences is more subtle – and possibly the already broad feature set warrants further extension to capture additional characteristics of more advanced levels.

For example, many of the syntactic features mentioned in the feature selection discussion

Feature	Group	Accur.
AoA	Psy	77%
IMAGERY	Psy	67.8%
CTTR	Lex	67.8%
MEANINGFULNESS	Psy	66.7%
CONCRETENESS	Psy	65.5%
FAMILIARITY	Psy	64.4%
MLC	Syn	64.4%
SUBTREES	Syn	64.4%
avg. senses per word	Lex	64.4%

Table 7: Accuracy of single feature ranking models for Int-Ele simplification

(SUBTREES, MLC, MLT) are correlated with text length. However, simplification can also involve sentence rewrites that do not affect the sentence length as such (e.g., paraphrasing, active/passive, reordering), which may warrant inclusion of features targeting more specific syntactic constructions or idiomatic word usage characteristic of advanced levels of English.

4.7 Error Analysis

Finally, to understand if there is a systematic pattern in the errors made by the ranker, we manually performed a qualitative analysis of errors. For this, we took the results of training with OSE3-Train data and testing with the OSE3-Test. This is the smallest test set (87 triplets), which given the 79.7% accuracy allows us to analyze 53 misclassified pairs. The following are four example sentence pairs/triplets from the test set. While the first two were ranked correctly by the ranker, the last two illustrate cases where the ranker failed.

Example 1

Adv: *He warned that it was too early to use oxytocin as a treatment for the social difficulties caused by autism and cautioned against buying oxytocin from suppliers online.*

Int: *He warned that it was too early to use oxytocin as a treatment for the social difficulties caused by autism and said people should not buy oxytocin online.*

Ele: *He said that it was too early to use oxytocin as a treatment for the social difficulties caused by autism and said people should not buy oxytocin online.*

Example 2

Int: *DNA taken from the wisdom tooth of a European hunter-gatherer has given scientists a glimpse of modern humans before the rise of farming.*

Ele: *Scientists have taken DNA from the tooth of a European hunter-gatherer and have found out what modern humans looked like before they started farming.*

Example 3

Adv: *Its inventor, Bob Propst, said in 1997, "the cubiclizing of people in modern corporations is monolithic insanity."*

Int: *Its inventor, Bob Propst, said, in 1997, "the use of cubicles in modern corporations is crazy."*

Ele: *The inventor, Bob Propst, said, in 1997, "the use of cubicles in modern companies is crazy."*

Example 4

Adv: *A special "auditor" declares him 96.9% "made in France" and Montebourg visits to present him with a medal.*

Int: *A special "auditor" declares him 96.9% "made in France" and Montebourg visited to present him with a medal.*

In Example 1, the transformation from *Adv* to *Int* is primarily paraphrasing ("and cautioned against buying oxytocin" vs. "and said people should not buy oxytocin") where was the transformation from *Int* to *Ele* is that of a simple lexical substitution ("He warned" vs. "He said"). However, in Example 2, there was more re-ordering and paraphrasing ("before the rise of farming" vs. "before they started farming"). In both these cases, our model correctly identified the changes as a simplification. The model thus effectively identifies paraphrases and lexical substitutions at multiple levels.

However, the model is not as effective for the sentence triplet in Example 3. It provides the correct pairwise rankings *Int* > *Ele* and *Adv* > *Ele*, but then incorrectly determines *Int* > *Adv*. So the model correctly identified a simple lexical substitution between *Int* and *Ele*, but failed to identify the transformation from “the cubiclizing of people” to “the use of cubicles” and from “monolithic insanity” to “crazy” as a simplification. This could possibly be because the parse structure as such did not alter much despite the rephrasing and neither “cubiclizing” nor the noun or lemmatized verb “cubiclize” exist in the psycholinguistic databases we used. Including broad coverage frequency measures of word usage could help address examples of this type. If the example at hand is typical, for the purpose of simplification evaluation at issue here word form frequencies would be preferable over lemma frequencies.

Finally, in Example 4 the only change between the two versions is a tense difference (“visits” vs. “visited”), which the model fails to rank correctly. It is debatable whether this change in tense indeed represents a simplification so that the case does not provide useful information on how to improve the approach.

Apart from the relevance of integrating broad-coverage frequency measures characterizing word form usage, our qualitative error analysis did not identify systematic failures of our models. The broad coverage of linguistic features integrated in a ranking approach successfully capture the relative differences in readability which characterize authentic simplification data at the sentence level.

5 Related Work

Evaluation of a text simplification approach is typically done in two ways. Most approaches are evaluated by comparing sentences using a combination of traditional readability measures, human fluency and grammaticality judgments of the generated output, and machine translation metrics (e.g., Barlacchi and Tonelli, 2013; Štajner et al., 2014; Siddharthan and Mandya, 2014). Some approaches evaluate the effect of text simplification on their target audience in terms of human recall and comprehension (e.g., Canning et al., 2000; Williams and Reiter, 2008; Bradley, 2012). Other recent work reported the usage of linguistic complexity measures that go beyond traditional readability formulae for evaluation of text simplification at a document level (Štajner and Saggion, 2013).

Comparing simplified versions of individual sentences with unsimplified versions in terms of text complexity is a rather recent endeavor. For example, sentence-level text complexity was explored in Intelligent Computer Assisted Language Learning to identify suitable sentences for creating learning exercises for German and Swedish learners (Segler, 2007; Pilán et al., 2014). Dell’Orletta et al. (2014) explored the linguistic nature of features contributing to sentential readability in the context of developing Italian text simplification system for adults with intellectual disabilities. However, the corpus used does not provide parallel texts with easy and difficult versions. In the absence of a sentence-aligned simplification corpus, the authors treat each sentence in the easy-to-read texts as easy. As Vajjala and Meurers (2014b, Fig. 1) showcases, this is a very problematic assumption. Even for a sentence-aligned simplification corpus such as the Wiki-SimpleWiki data set the only thing guaranteed is that there is one sentence which is harder than the simple one.

Napoles and Dredze (2010) considered a binary classification of Wiki-SimpleWiki at both text and sentence level, using a range of lexical and syntactic features. They also work with the simplifying assumption that all sentences in Wikipedia are difficult and those in SimpleWiki are simple. An interesting aspect of the results of Napoles and Dredze (2010) and also of Dell’Orletta et al. (2014) is that they achieve the highest classification accuracy at text and sentence level when combining all features.

Vajjala and Meurers (2014a) compared sentences in terms of their reading levels using their readability model and showed that sophisticated linguistically motivated readability models can effectively identify the differences between sentences. In the current paper, we extend their research by exploring sentential simplification evaluation as a ranking problem and showed that ranking achieves superior performance to regression for this task.

Ranking has been used for readability assessment at the document level (Ma et al., 2012) and for related tasks such as ordering MT system output sentences in terms of their language quality (Avramidis, 2013), and for ranking sentences in opinion summarization (Kim et al., 2013). To the best of our knowledge, simplification evaluation was not explored as a ranking problem before.

6 Conclusions

We presented an approach to evaluate automatic text simplification systems in terms of the reading level of individual sentences. The approach involves the use of a pairwise ranking approach to compare unsimplified and simplified versions of sentences in terms of the reading level. It identifies the order in terms of their reading level correctly with an accuracy of over 80%, the best accuracy for this task we are aware of. We performed in-corpus and cross-corpus evaluations with two very different sentence-aligned corpora and showed that the approach generalizes well across corpora. The approach performs at a level that should make it useful in practice to automatically evaluate text simplification for language learners in real-life educational settings.

In terms of the linguistic nature of simplification, we studied the role of individual features and groups of features in predicting the ranking order between simplified and unsimplified versions of the sentences. We found that for the OSE data set, psycholinguistic features such as Age-of-Acquisition are the most predictive individual features. For the Wiki-SimpleWiki data set, syntactic features also figure prominently. However, an approach using the full range of features systematically results in the best performance and generalizes best in the cross-corpus settings.

Pursuing the question whether there is a singular notion of simplification, we studied the differences in text simplification occurring at different levels of readability. Our features identify simplification between intermediate and elementary levels better compared than between advanced and intermediate level. It is possible that this is due to a higher degree of simplification between the former, but we also plan to study whether other types of features could be added to identify characteristics at higher levels of readability, such as features targeting specific constructions or idiomatic usage. The small qualitative error analysis we performed revealed that a broad coverage method capturing the frequency of word usage may further improve results.

To carry out the research, we created a new corpus of sentence-aligned simplified texts based on OneStopEnglish texts rewritten by experts for language learners into three reading levels. The new corpus resource can empirically enrich future research on sentence-level simplification, helping to ensure that the results obtained are more generally valid than for the single Wiki-SimpleWiki resource that was available so far.

Outlook Adding frequency features capturing word usage and exploring feature selection in more detail by selecting the best features for the ranker while removing the correlated ones (Geng et al., 2007) are the immediate directions we would like to pursue.

It would also be interesting to apply the approach to evaluate the output of automatic text simplification systems and compare their performance in terms of readability. Going beyond complexity, in the long term it could be interesting to extend the approach to a full framework for evaluating automatic text simplification systems by integrating aspects of fluency and grammaticality.

References

Eleftherios Avramidis. 2013. Sentence-level ranking with quality estimation. *Machine Translation*, 27(3-4):239–256.

R. H. Baayen, R. Piepenbrock, and L. Gulikers. 1995. The CELEX lexical databases. CDROM, `http://www.ldc.upenn.edu/Catalog/readme_files/celex.readme.html`.

Gianni Barlacchi and Sara Tonelli. 2013. Ernesta: A sentence simplification tool for children's stories in italian. In *14th International Conference on Computational Linguistics and Intelligent Text Processing, (CICLing)*, pages 476–487. Springer.

Jeremy Bradley. 2012. *Computergesteuerte Hilfe für deutschsprachige Aphasiker und Aphasikerinnen*. Ph.D. thesis, Technische Universität Wien, Wien.

C.J.C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the International Conference on Machine Learning*, pages 89–96.

Arnaldo Candido, Jr., Erick Maziero, Caroline Gasperin, Thiago A. S. Pardo, Lucia Specia, and Sandra M. Aluisio. 2009. Supporting the adaptation of texts for poor literacy readers: a text simplification editor for brazilian portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, EdAppsNLP '09, pages 34–42, Stroudsburg, PA, USA.

Yvonne Canning, John Tait, Jackie Archibald, and Ros Crawley. 2000. Cohesive generation of syntactically simplified newspaper text. In *Third International Workshop on Text, Speech and Dialogue, TSD 2000, Brno, Czech Republic, September 13-16, 2000*, pages 145–150. Springer.

CCSSO. 2010. Common core state standards for english language arts & literacy in history/social studies, science, and technical subjects. appendix B: Text exemplars and sample performance tasks. Technical report, National Governors Association Center for Best Practices, Council of Chief State School Officers. http://www.corestandards.org/assets/Appendix_B.pdf.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, pages 1041–1044.

Felice Dell'Orletta, Martijn Wieling, Andrea Cimino, Giulia Venturi, and Simonetta Montemagni. 2014. Assessing the readability of sentences: Which corpora and features? In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications (BEA9)*, pages 163–173, Baltimore, Maryland, USA. ACL.

William H. DuBay. 2006. *The Classic Readability Studies*. Impact Information, Costa Mesa, California.

Y. Freund, R. Iyer, R. Schapire, , and Y. Singer. 2003. An efficient boosting algorithm for combining preferences. *The Journal of Machine Learning Research*, 4:933–969.

Xiubo Geng, Tie-Yan Liu, Tao Qin, and Hang Li. 2007. Feature selection for ranking. In *Proceedings of SIGIR Conference*, pages 548–552.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer, 2000. *Large margin rank boundaries for ordinal regression*, pages 115–132. MIT Press, Cambridge, MA.

Sujay Kumar Jauhar and Lucia Specia. 2012. Uowshef: Simplex – lexical simplicity ranking based on contextual and psycholinguistic features. In *In proceedings of the First Joint Conference on Lexical and Computational Semantics (SEM)*.

T. Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–226. ACM Press.

Siddhartha Jonnalagadda and Graciela Gonzalez. 2010. Biosimplify: an open source sentence simplification engine to improve recall in automatic biomedical information extraction. In *AMIA Annual Symposium Proceedings*, pages 351–356.

Hyun Duk Kim, Malu G Castellanos, Meichun Hsu, ChengXiang Zhai, Umeshwar Dayal, and Riddhiman Ghosh. 2013. Ranking explanatory sentences for opinion summarization. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 1069–1072.

George R. Klare. 1974. Assessing readability. *Reading Research Quarterly*, 10(1):62–102.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4):978–990.

Hang Li. 2014. *Learning to Rank for Information Retrieval and Natural Language Processing*. Morgan and Claypool Publishers.

Yi Ma, Eric Fosler-Lussier, and Robert Lofthus. 2012. Ranking-based readability assessment for early primary children's literature. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 548–552, Stroudsburg, PA, USA. Association for Computational Linguistics.

Julie Medero and Marie Ostendorf. 2011. Identifying targets for syntactic simplification. In *ISCA International Workshop on Speech and Language Technology in Education (SLaTE 2011)*.

George Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, November.

Courtney Napoles and Mark Dredze. 2010. Learning simple wikipedia: a cogitation in ascertaining abecedarian language. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, CL&W '10, pages 42–50, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rani Nelken and Stuart M. Shieber. 2006. Towards robust context-sensitive sentence alignment for monolingual corpora. In *In 11th Conference of the European Chapter of the Association of Computational Linguistics*, pages 161–168. Assoc. for Computational Linguistics.

J. Nelson, C. Perfetti, D. Liben, and M. Liben. 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. Technical report, The Council of Chief State School Officers.

Sarah E. Petersen and Mari Ostendorf. 2007. Text simplification for language learners: A corpus analysis. In *Speech and Language Technology for Education (SLaTE)*, pages 69–72.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. Rule-based and machine learning approaches for second language sentence-level readability. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications (BEA9)*, pages 174–184, Baltimore, Maryland, USA. ACL.

Thomas M. Segler. 2007. *Investigating the Selection of Example Sentences for Unknown Target Words in ICALL Reading Texts for L2 German*. Ph.D. thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.

Advaith Siddharthan and Angrosh Mandya. 2014. Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 722–731, Gothenburg, Sweden, April. ACL.

Advaith Siddharthan. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics: Special issue on Recent Advances in Automatic Readability Assessment and Text Simplification*, 165:2:259–298.

Gokhan Tur, Dilek Hakkani-Tür, Larry Heck, and S. Parthasarathy. 2011. Sentence simplification for spoken language understanding. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP)*, pages 5628–5631.

Sowmya Vajjala and Detmar Meurers. 2014a. Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 288–297, Gothenburg, Sweden, April. ACL, Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2014b. Readability assessment for text simplification: From analyzing documents to identifying sentential simplifications. *International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification*.

Sandra Williams and Ehud Reiter. 2008. Generating basic skills reports for low-skilled readers*. *Nat. Lang. Eng.*, 14:495–525, October.

M.D. Wilson. 1988. The MRC psycholinguistic database: Machine readable dictionary, version 2. *Behavioural Research Methods, Instruments and Computers*, 20(1):6–11.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 409–420. Assoc. for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of The 23rd International Conference on Computational Linguistics (COLING), August 2010. Beijing, China*, pages 1353–1361.

Sanja Štajner and Horacio Saggion. 2013. Readability indices for automatic evaluation of text simplification systems: A feasibility study for spanish. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 374–382, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Sanja Štajner, Ruslan Mitkov, and Horacio Saggion. 2014. One step closer to automatic evaluation of text simplification systems. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 1–10, Gothenburg, Sweden. ACL.