Assessing the relative reading level of sentence pairs for text simplification

Sowmya Vajjala and Detmar Meurers
LEAD Graduate School, Seminar für Sprachwissenschaft
Universität Tübingen
{sowmya,dm}@sfs.uni-tuebingen.de

Abstract

While the automatic analysis of the readability of texts has a long history, the use of readability assessment for text simplification has received only little attention so far. In this paper, we explore readability models for identifying differences in the reading levels of simplified and unsimplified versions of sentences.

Our experiments show that a relative ranking is preferable to an absolute binary one and that the accuracy of identifying relative simplification depends on the initial reading level of the unsimplified version. The approach is particularly successful in classifying the relative reading level of harder sentences.

In terms of practical relevance, the approach promises to be useful for identifying particularly relevant targets for simplification and to evaluate simplifications given specific readability constraints.

1 Introduction

Text simplification essentially is the process of rewriting a given text to make it easier to process for a given audience. The target audience can either be human users trying to understand a text or machine applications, such as a parser analyzing text. Text simplification has been used in a variety of application scenarios, from providing simplified newspaper texts for aphasic readers (Canning and Tait, 1999) to supporting the extraction of protein-protein interactions in the biomedical domain (Jonnalagadda and Gonzalez, 2009).

A related field of research is automatic readability assessment, which can be useful for evaluating text simplification. It can also be relevant for intermediate simplification steps, such as the identification of target sentences for simplification. Yet, so far there has only been little research connecting the two subfields, possibly because readability research typically analyzes documents, whereas simplification approaches generally targeted lexical and syntactic aspects at the sentence level. In this paper, we attempt to bridge this gap between readability and simplification by studying readability at a sentence level and exploring how well can a readability model identify the differences between unsimplified and simplified sentences.

Our main research questions in this paper are: 1. Can the readability features that worked at the document level successfully be used at the sentence level? 2. How accurately can we identify the differences in the sentential reading level before and after simplification? To pursue these questions, we started with constructing a document-level readability model. We then applied it to normal and simplified versions of sentences drawn from Wikipedia and Simple Wikipedia.

As context of our work, we first discuss relevant related research. Section 2 then describes the corpora and the features we used to construct our readability model. Section 3 discusses the performance of our readability model in comparison with other existing systems. Sections 4 and 5 present our experiments with sentence level readability analysis and the results. In Section 6 we present our conclusions and plans for future work.

1.1 Related Work

Research into automatic text simplification essentially started with the idea of splitting long sentences into multiple shorter sentences to improve parsing efficiency (Chandrasekar et al., 1996; Chandrasekar and Srinivas, 1996). This was followed by rule-based approaches targeting human and machine uses (Carroll et al., 1999; Siddharthan, 2002, 2004).

With the availability of a sentence-aligned corpus based on Wikipedia and SimpleWikipedia

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 288–297,
Gothenburg, Sweden, April 26-30 2014.

texts, data-driven approaches, partly inspired by statistical machine translation, appeared (Specia, 2010; Zhu et al., 2010; Bach et al., 2011; Coster and Kauchak, 2011; Woodsend and Lapata, 2011).

While simplification methods have evolved, understanding which parts of a text need to be simplified and methods for evaluating the simplified text so far received only little attention. The use of readability assessment for simplification has mostly been restricted to using traditional readability formulae for evaluating or generating simplified text (Zhu et al., 2010; Wubben et al., 2012; Klerke and Søgaard, 2013; Stymne et al., 2013). Some recent work briefly addresses issues such as classifying sentences by their reading level (Napoles and Dredze, 2010) and identifying sentential transformations needed for text simplification using text complexity features (Medero and Ostendorf, 2011). Some simplification approaches for non-English languages (Aluisio et al., 2010; Gasperin et al., 2009; Štajner et al., 2013) also touch on the use of readability assessment.

In the present paper, we focus on the neglected connection between readability analysis and simplification. We show through a cross-corpus evaluation that a document level, regression-based readability model successfully identifies the differences between simplified vs. unsimplified sentences. This approach can be useful in various stages of simplification ranging from identifying simplification targets to the evaluation of simplification outcomes.

2 Corpora and Features

2.1 Corpora

We built and tested our document and sentence level readability models using three publicly available text corpora with reading level annotations.

WeeBit Corpus: The WeeBit corpus (Vajjala and Meurers, 2012) consists of 3,125 articles belonging to five reading levels, with 625 articles per reading level. The texts compiled from the WeeklyReader and BBC Bitesize target English language learners from 7 to 16 years of age. We used this corpus to build our primary readability model by mapping the five reading levels in the corpus to a scale of 1–5 and considered readability assessment as a regression problem.

Common Core Standards Corpus: This corpus consists of 168 English texts available from the Appendix B of the Common Core Standards reading initiative of the U.S. education system (CCSSO, 2010). They are annotated by experts with grade bands that cover the grades 1 to 12. These texts serve as exemplars for the level of reading ability at a given grade level. This corpus was introduced as an evaluation corpus for readability models in the recent past (Sheehan et al., 2010; Nelson et al., 2012; Flor et al., 2013), so we used it to compare our model with other systems.

Wiki-SimpleWiki Sentence Aligned Corpus: This corpus was created by Zhu et al. (2010) and consists of $\sim$100k aligned sentence pairs drawn from Wikipedia and Simple English Wikipedia. We removed all pairs of identical sentences, i.e., where the Wiki and the SimpleWiki versions are the same. We used this corpus to study reading level assessment at the sentence level.

2.2 Features

We started with the feature set described in Vajjala and Meurers (2012) and added new features focusing on the morphological and psycholinguistic properties of words. The features can be broadly classified into four groups.

Lexical richness and POS features: We adapted the lexical features from Vajjala and Meurers (2012). This includes measures of lexical richness from Second Language Acquisition (SLA) research and measures of lexical variation (noun, verb, adjective, adverb and modifier variation). In addition, this feature set also includes part-of-speech densities (e.g., the average # of nouns per sentence). The information needed to calculate these features was extracted using the Stanford Tagger (Toutanova et al., 2003). None of the lexical richness and POS features we used refer to specific words or lemmas.

Syntactic Complexity features: Parse tree based features and some syntactic complexity measures derived from SLA research proved useful for readability classification in the past, so we made use of all the syntactic features from Vajjala and Meurers (2012): mean lengths of various production units (sentence, clause, t-unit), measures of coordination and subordination (e.g., # of coordinate clauses per clause), the presence of particular syntactic structures (e.g., VPs per t-unit), the number of phrases of various categories (e.g., NP, VP, PP), the average lengths

of phrases, the parse tree height, and the number of constituents per subtree. None of the syntactic features refer to specific words or lemmas. We used the BerkeleyParser (Petrov and Klein, 2007) for generating the parse trees and the Tregex tool (Levy and Andrew, 2006) to count the occurrences of the syntactic patterns.

While the first two feature sets are based on our previous work, as far as we know the next two are used in readability assessment for the first time.

Features from the Celex Lexical Database: The Celex Lexical Database (Baayen et al., 1995) is a database consisting of information about morphological, syntactic, orthographic and phonological properties of words along with word frequencies in various corpora. Celex for English contains this information for more than 50,000 lemmas. An overview of the fields in the Celex database is provided online[1] and the Celex user manual[2].

We used the morphological and syntactic properties of lemmas as features. We excluded word frequency statistics and properties which consisted of word strings. In all, we used 35 morphological and 49 syntactic properties that were expressed using either character or numeric codes in this database as features for our task.

The morphological properties in Celex include information about the derivational, inflectional and compositional features of the words, their morphological origins and complexity. The syntactic properties of the words in Celex describe the attributes of a word depending on its parts of speech. For the morphological and syntactic properties from this database, we used the proportion of occurrences per text as features. For example, the ratio of transitive verbs, complex morphological words, and vocative nouns to number of words. Lemmas from the text that do not have entries in the Celex database were ignored.

Word frequency statistics from Celex have been used before to analyze text difficulty in the past (Crossley et al., 2007). However, to our knowledge, this is the first time morphological and syntactic information from the Celex database is used for readability assessment.

Psycholinguistic features: The MRC Psycholinguistic Database (Wilson, 1988) is a freely available, machine readable dictionary annotated with 26 linguistic and psychological attributes of about 1.5 million words.[3] We used the measures of word familiarity, concreteness, imageability, meaningfulness, and age of acquisition from this database as our features, by encoding their average values per text.

Kuperman et al. (2012) compiled a freely available database that includes Age of Acquisition (AoA) ratings for over 50,000 English words.[4] This database was created through crowd sourcing and was compared with several other AoA norms, which are also included in the database. For each of the five AoA norms, we computed the average AoA of words per text.

Turning to the final resource used, we included the average number of senses per word as calculated using the MIT Java WordNet Interface as a feature.[5] We excluded auxiliary verbs for this calculation as they tend to have multiple senses that do not necessarily contribute to reading difficulty.

Combining the four feature groups, we encode 151 features for each text.

3 Document-Level Readability Model

In our first experiment, we tested the document-level readability model based on the 151 features using the WeeBit corpus. Under a regression perspective on readability, we evaluated the approach using Pearson Correlation and Root Mean Square Error (RMSE) in a 10-fold cross-validation setting. We used the SMO Regression implementation from WEKA (Hall et al., 2009) and achieved a Pearson correlation of 0.92 and an RMSE of 0.53.

The document-level performance of our 151 feature model is virtually identical to that of the regression model we presented in Vajjala and Meurers (2013). But compared to our previous work, the Celex and psycholinguistic features we included here provide more lexical information that is meaningful to compute even for the sentence-level analysis we turn to in the next section.

To be able to compare our document-level results with other contemporary readability approaches, we need a common test corpus. Nelson et al. (2012) compared several state of the art readability assessment systems using five test sets and showed that the systems that went beyond traditional formulae and wordlists performed better

[1] http://celex.mpi.nl/help/elemmas.html

[2] http://catalog.ldc.upenn.edu/docs/LDC96L14

[3] http://www.psych.rl.ac.uk

[4] http://crr.ugent.be/archives/806

[5] http://projects.csail.mit.edu/jwi

on these real-life test sets. We tested our model on one of the publicly accessible test corpora from this study, the Common Core Standards Corpus. Flor et al. (2013) used the same test set to study a measure of lexical tightness, providing a further performance reference.

Table 1 compares the performance of our model to that reported for several commercial (indicated in italics) and research systems on this test set. Nelson et al. (2012) used Spearman's Rank Correlation and Flor et al. (2013) used Pearson Correlation as evaluation metrics. To facilitate comparison, for our approach we provide both measures.

System	Spearman	Pearson
Our System	**0.69**	**0.61**
Nelson et al. (2012):		
REAP[6]	0.54	–
ATOS[7]	0.59	–
DRP[8]	0.53	–
Lexile[9]	0.50	–
Reading Maturity[10]	**0.69**	–
SourceRater[11]	**0.75**	–
Flor et al. (2013):		
Lexical Tightness	–	-0.44
Flesch-Kincaid	–	0.49
Text length	–	0.36

Table 1: Performance on CommonCore data

As the table shows, our model is the best non-commercial system and overall second (tied with the Reading Maturity system) to SourceRater as the best performing commercial system on this test set. These results on an independent test set confirm the validity of our document-level readability model. With this baseline, we turned to a sentence-level readability analysis.

4 Sentence-Level Binary Classification

For each of the pairs in the Wiki-SimpleWiki Sentence Aligned Corpus introduced above, we labeled the sentence from Wikipedia as *hard* and that from Simple English Wikipedia as *simple*. The corpus thus consisted of single sentences, each labeled either *simple* or *hard*. On this basis, we constructed a binary classification model.

[6] http://reap.cs.cmu.edu
[7] http://renlearn.com/atos
[8] http://questarai.com/Products/DRPProgram
[9] http://lexile.com
[10] http://readingmaturity.com
[11] http://naeptba.ets.org/SourceRater3

Our document-level readability model does not include discourse features, so all 151 features can also be computed for individual sentences. We built a binary sentence-level classification model using WEKA's Sequential Minimal Optimization (SMO) for training an SVM in WEKA on the Wiki-SimpleWiki sentence aligned corpus. The choice of algorithm was primarily motivated by the fact that it was shown to be efficient in previous work on readability classification (Feng, 2010; Hancke et al., 2012; Falkenjack et al., 2013).

The accuracy of the resulting classifier determining whether a given sentence is *simple* or *hard* was disappointing, reaching only 66% accuracy in a 10-fold cross-validation setting. Experiments with different classification algorithms did not yield any more promising results. To study how the classification performance is impacted by the size of the training data, we experimented with different sizes, using SMO as the classification algorithm. Figure 1 shows the classification accuracy with different training set sizes.

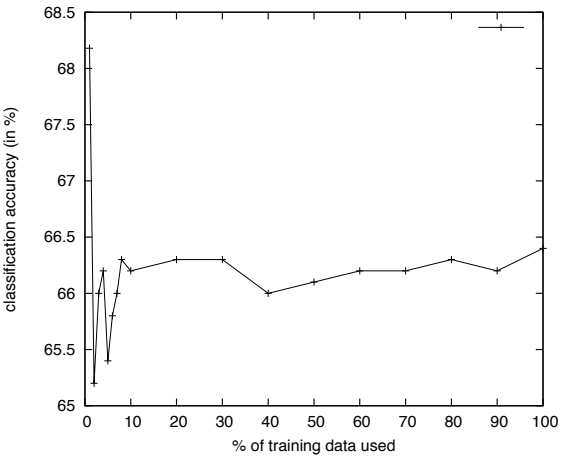

Figure 1: Training size vs. classification accuracy

The graph shows that beyond 10% of the training data, more training data did not result in significant differences in classification accuracy. Even at 10%, the training set contains around 10k instances per category, so the variability of any of the patterns distinguished by our features is sufficiently represented.

We also explored whether feature selection could be useful. A subset of features chosen by removing correlated features using the CfsSubsetEval method in WEKA did not improve the results, yielding an accuracy of 65.8%. A simple baseline based on the sentence length as single feature results in an accuracy of 60.5%, underscoring the

limited value of the rich feature set in this binary classification setup.

For the sake of a direct comparison with the document-level model, we also explored modeling the task as a regression on a 1–2 scale. In comparison to the document-level model, which as discussed in section 3 had a correlation of 0.92, the sentence-level model achieves only a correlation of 0.4. A direct comparison is also possible when we train the document-level model as a five-class classifier with SMO. This model achieved a classification accuracy of $\sim$90% on the documents, compared to the 66% accuracy of the sentence-level model classifying sentences. So under each of these perspectives, the sentence-level models on the sentence task are much less successful than the document-level models on the document task.

But does this indicate that it is not possible to accurately identify the reading level distinctions between simplified and unsimplified versions at the sentence level? Is there not enough information available when considering a single sentence?

We hypothesized that the drop in the classification accuracy instead results from the relative nature of simplification. For each pair of the Wiki-SimpleWiki sentence aligned corpus we used, the Wiki sentence was harder than the SimpleWikipedia sentence. But this does not necessarily mean that each of the Wikipedia sentences is harder than each of the SimpleWikipedia sentences. The low accuracy of the binary classifier may thus simply result from the inappropriate assumption of an absolute, binary classification viewing each of the sentences originating from SimpleWikipedia as simple and each from the regular Wiki as hard.

The confusion matrices of the binary classification suggests some support for this hypothesis, as more *simple* sentences were classified as *hard* compared to the other way around. This can result when a *simple* sentence is simpler than its *hard* version, but could actually be simplified further – and as such may still be harder than another unsimplified sentence. The hypothesis thus amounts to saying that the two-class classification model mistakenly turned the relative difference between the sentence pairs into a global classification of individual sentences, independent of the pairs they occur in.

How can we verify this hypothesis? The sentence corpus only provides the relative ranking of the pairs, but we can try to identify more fine-grained readability levels for sentences by applying the five class readability model for documents that was introduced in section 3.

5 Relative Reading Levels of Sentences

We applied the document-level readability model to the individual sentences from the Wiki-SimpleWiki corpus to study which reading levels are identified by our model. As we are using a regression model, the values sometimes go beyond the training corpus' scale of 1–5. For ease of comparison, we rounded off the reading levels to the five level scale, i.e., 1 means 1 or below, and 5 means 5 or above. Figure 2 shows the distribution of Wikipedia and SimpleWikipedia sentences according to the predictions of our document-level readability model trained on the WeeBit corpus.

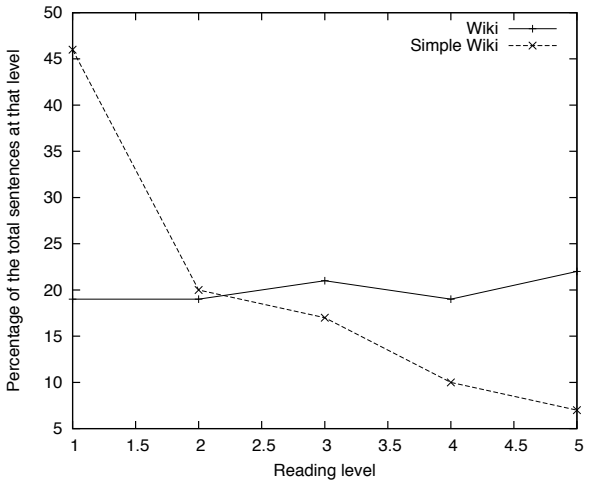

Figure 2: Reading level distribution of the Wikipedia and SimpleWikipedia sentences

The model determines that a high percentage of the SimpleWiki sentences belong to lower reading levels, with over 45% at the lowest reading level; yet there also are some SimpleWikipedia sentences which are aligned even to the highest readability level. In contrast, the regular Wikipedia sentences are evenly distributed across all reading levels.

The distributions identified by the model support our hypothesis that some Wiki sentences are simpler than some SimpleWikipedia sentences. Note that this is fully compatible with the fact that for each pair of (SimpleWiki,Wiki) sentences included in the corpus, the former is higher in reading level than the latter; e.g., just consider two sentence pairs with the levels $(1, 2)$ and $(3, 5)$.

5.1 On the discriminating power of the model

Zooming in on the relative reading levels of the paired unsimplified and simplified sentences, we wanted to determine for how many sentence pairs the sentence reading levels determined by our model are compatible with the pair's ranking. In other words, we calculated the percentage of pairs (S, N) in which the reading level of a simplified sentence (S) is identified as less than, equal to, or greater than the unsimplified (normal) version of the sentence (N), i.e., $S < N$, $S = N$, and $S > N$. Where simplification split a sentence into multiple sentences, we computed S as the average reading level of the split sentences.

Given the regression model setup, we can consider how big the difference between two reading levels determined by the model should be in order for us to interpret it as a categorical difference in reading level. Let us call this discriminating reading-level difference the *d-level*. For example, with $d = 0.3$, a sentence pair determined to be at levels $(3.4, 3.2)$ would be considered a case of $S = N$, whereas $(3.4, 3.7)$ would be an instance of $S < N$. The d-value can be understood as a measure of how fine-grained the model is in identifying reading-level differences between sentences.

If we consider the percentage of samples identified as $S <= N$ as an accuracy measure, Figure 3 shows the accuracy for different d-values.

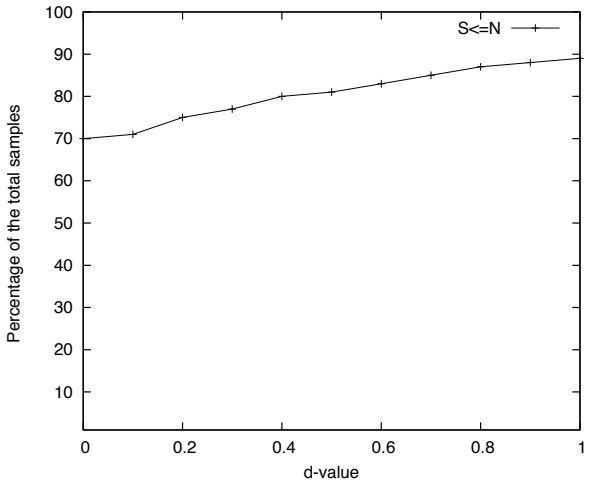

Figure 3: Accurately identified $S <= N$

We can observe that the percentage of instances that the model correctly identifies as $S <= N$ steadily increases from 70% to 90% as *d* increases. While the value of d in theory can be anything, values beyond 1 are uninteresting in the context of this study. At $d = 1$, most of the sentence pairs already belong to $S = N$, so increasing this further would defeat the purpose of identifying reading-level differences. The higher the d-value, the more of the simplified and unsimplified pairs are lumped together as indistinguishable.

Spelling out the different cases from Figure 3, the number of pairs identified correctly, equated, and misclassified as a function of the d-value is shown in Figure 4.

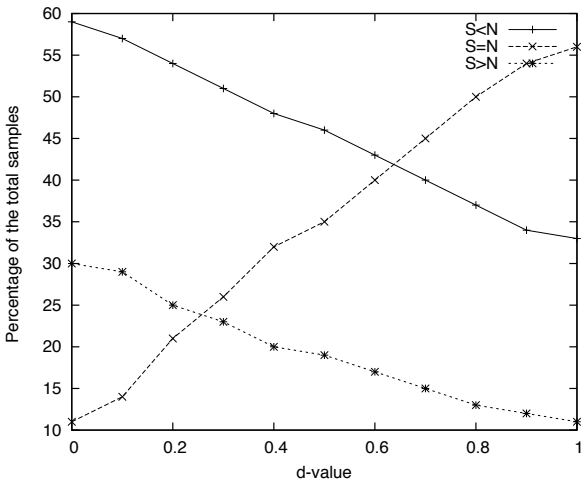

Figure 4: Correctly ($S < N$), equated ($S = N$), and incorrectly ($S > N$) identified sentence pairs

At $d = 0.4$, around 50% of the pairs are correctly classified, 20% are misclassified, and 30% equated. At $d = 0.7$, the rate of pairs for which no distinction can be determined already rises above 50%. For d-values between 0.3 and 0.6, the percentage of correctly identified pairs exceeds the percentage of equated pairs, which in turn exceeds the percentage of misclassified pairs.

5.2 Influence of reading-level on accuracy

We saw in Figure 2 that the Wikipedia sentences are uniformly distributed across the reading levels, and for each of these sentences, a human simplified version is included in the corpus. Even sentences identified by our readability model as belonging to the lower reading levels thus were further simplified. This leads us to investigate whether the reading level of the unsimplified sentence influences the ability of our model to correctly identify the simplification relationship.

To investigate this, we separately analyzed pairs where the unsimplified sentences had a higher reading level and those where it had a lower reading level, taking the middle of the scale (2.5) as the

cut-off point. Figure 5 shows the accuracies obtained when distinguishing unsimplified sentences of two readability levels.

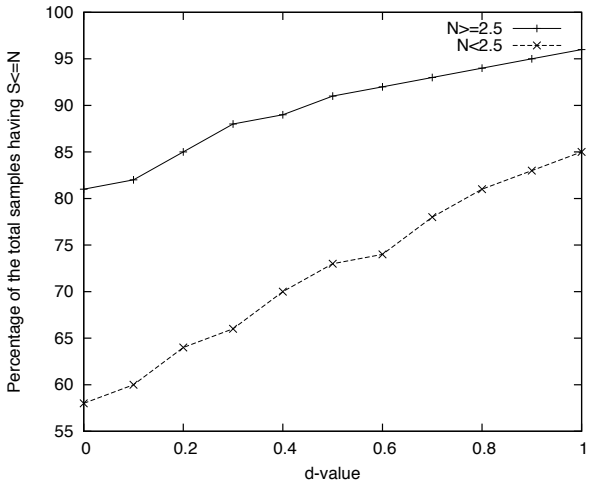

Figure 5: Accuracy ($S <= N$) for different N types

For the pairs where the reading level of the unsimplified version is high, the accuracy of the readability model is high (80–95%). In the other case, the accuracy drops to 65–75% (for $0.3 <= d <= 0.6$). Presumably the complex sentences for which the model performs best offer more syntactic and lexical material informing the features used.

When we split the graph into the three cases again ($S < N$, $S = N$, $S > N$), the pairs with a high-level unsimplified sentence in Figure 6 follow the overall picture of Figure 4.

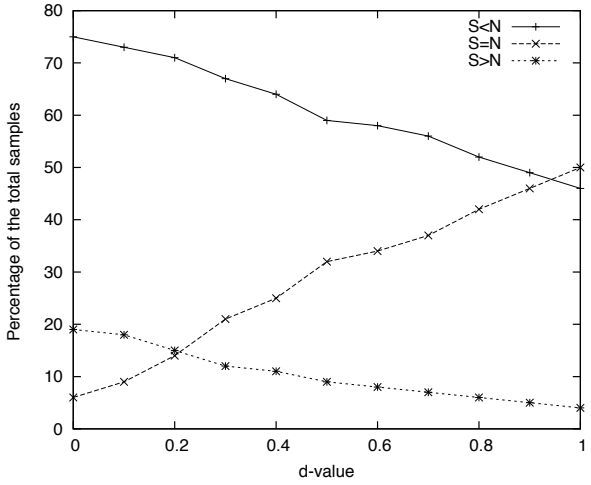

Figure 6: Results for $N >= 2.5$

On the other hand, the results in Figure 7 for the pairs with an unsimplified sentence at a low readability level establish that the model essentially is incapable to identify readability differences.

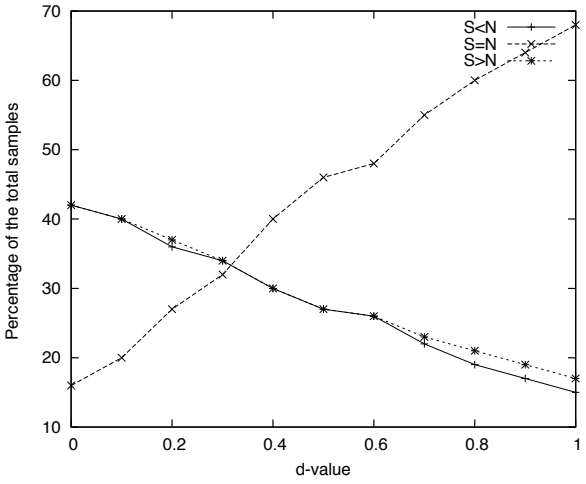

Figure 7: Results for $N < 2.5$

The correctly identified $S < N$ and the incorrectly identified $S > N$ cases mostly overlap, indicating chance-level performance. Increasing the d-level only increases the number of equated pairs, without much impact on the number of correctly distinguished pairs.

In real-world terms, this means that it is difficult to identify simplifications of an already simple sentence. While some of this difficulty may stem from the fact that simple sentences are likely to be shorter and thus offer less linguistic material on which an analysis can be based, it also points to a need for more research on features that can reliably distinguish lower levels of readability.

Summing up, the experiments discussed in this section show that a document-level readability model trained on the WeeBit corpus can provide insightful perspectives on the nature of simplification at the sentence level. The results emphasize the relative nature of readability and the need for more features capable of identifying characteristics distinguishing sentences at lower levels.

6 Conclusions

We started with constructing a document-level readability model and compared its performance with other readability systems on a standard test set. Having established the state-of-the-art performance of our document-level model, we moved on to investigate the use of the features and the model at the sentence level.

In the sentence-level research, we first used the same feature set to construct a two-class readability model on the sentences from the Wikipedia-SimpleWikipedia sentence aligned corpus. The

model only achieved a classification accuracy of 66%. Exploring the causes for this low performance, we studied the sentences in the aligned pairs through the lens of our document-level readability model, the regression model based on the five level data of the WeeBit corpus. Our experiment identifies most of the Simple Wikipedia sentences as belonging to the lower levels, with some sentences also showing up at higher levels. The sentences from the normal Wikipedia, on the other hand, display a uniform distribution across all reading levels. A simplified sentence (S) can thus be at a lower reading level than its paired unsimplified sentence (N) while also being at a higher reading level than another unsimplified sentence. Given this distribution of reading levels, the low performance of the binary classifier is expected. Instead of an absolute, binary difference in reading levels that counts each Wikipedia sentence from the corpus as hard and each Simple Wikipedia sentence as simple, a relative ranking of reading levels seems to better suit the data.

Inspecting the relative difference in the reading levels of the aligned unsimplified-simplified sentence pairs, we characterized the accuracy of predicting the relative reading level ranking in a pair correctly depending on the reading-level difference d required to required to identify a categorical difference. While the experiments were performed to verify the hypothesis that simplification is relative, they also confirm that the document-level readability model trained on the WeeBit corpus generalized well to Wikipedia-SimpleWikipedia as a different, sentence-level corpus.

The analysis revealed that the accuracy depends on the initial reading level of the unsimplified sentence. The model performs very well when the reading level of the unsimplified sentence is higher, but the features seem limited in their ability to pick up on the differences between sentences at the lowest levels. In future work, we thus intend to add more features identifying differences between lower levels of readability.

Taking the focus on the relative ranking of the readability of sentences one step further, we are currently studying if modeling the readability problem as preference learning or ordinal regression will improve the accuracy in predicting the relation between simplified and unsimplified sentence versions.

Overall, the paper contributes to the state of the art by providing a methodology to quantitatively evaluate the degree of simplification performed by an automatic system. The results can also be potentially useful in providing assistive feedback for human writers preparing simplified texts given specific target user constraints. We plan to explore the idea of generating simplified text with readability constraints as suggested in Stymne et al. (2013) for Machine Translation.

Acknowledgements

We thank the anonymous reviewers for their detailed comments. Our research was funded by the LEAD Graduate School (GSC 1028, `http://purl.org/lead`), a project of the Excellence Initiative of the German federal and state governments, and the European Commission's 7th Framework Program under grant agreement number 238405 (CLARA).

References

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9.

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`.

Nguyen Bach, Qin Gao, Stephan Vogel, and Alex Waibel. 2011. Tris: A statistical sentence simplifier with log-linear models and margin-based discriminative training. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 474–482. Asian Federation of Natural Language Processing.

Yvonne Canning and John Tait. 1999. Syntactic simplification of newspaper text for aphasic readers. In *Proceedings of SIGIR-99 Workshop on Customised Information Delivery*, pages 6–11.

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 269–270.

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 and B. Srinivas. 1996. Automatic induction of rules for text simplification. Technical Report IRCS Report 96–30, Upenn, NSF Science and Technology Center for Research in Cognitive Science.

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.

William Coster and David Kauchak. 2011. Simple english wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA, June. Association for Computational Linguistics.

Scott A. Crossley, David F. Dufty, Philip M. McCarthy, and Danielle S. McNamara. 2007. Toward a new readability: A mixed model approach. In Danielle S. McNamara and Greg Trafton, editors, *Proceedings of the 29th annual conference of the Cognitive Science Society*. Cognitive Science Society.

Johan Falkenjack, Katarina Heimann Mühlenbock, and Arne Jönsson. 2013. Features indicating readability in swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*.

Lijun Feng. 2010. *Automatic Readability Assessment*. Ph.D. thesis, City University of New York (CUNY).

Michael Flor, Beata Beigman Klebanov, and Kathleen M. Sheehan. 2013. Lexical tightness and text complexity. In *Proceedings of the Second Workshop on Natural Language Processing for Improving Textual Accessibility*.

Caroline Gasperin, Lucia Specia, Tiago F. Pereira, and Sandra M. Aluisio. 2009. Learning when to simplify sentences for natural text simplification. In *Encontro Nacional de Inteligência Artificial (ENIA-2009)*.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. In *The SIGKDD Explorations*, volume 11, pages 10–18.

Julia Hancke, Detmar Meurers, and Sowmya Vajjala. 2012. Readability classification for german using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1063–1080, Mumbay, India.

Siddhartha Jonnalagadda and Graciela Gonzalez. 2009. Sentence simplification aids protein-protein interaction extraction. In *Proceedings of The 3rd International Symposium on Languages in Biology and Medicine, Jeju Island, South Korea, November 8-10, 2009*.

Sigrid Klerke and Anders Søgaard. 2013. Simple, readable sub-sentences. In *Proceedings of the ACL Student Research Workshop*.

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.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

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

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.

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.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York, April.

Kathleen M. Sheehan, Irene Kostin, Yoko Futagi, and Michael Flor. 2010. Generating automated text complexity classifications that are aligned with targeted text complexity standards. Technical Report RR-10-28, ETS, December.

Advaith Siddharthan. 2002. An architecture for a text simplification system. In *In Proceedings of the Language Engineering Conference 2002 (LEC 2002)*.

Advaith Siddharthan. 2004. Syntactic simplification and text cohesion. Technical Report UCAM-CL-TR-597, University of Cambridge Computer Laboratory.

Lucia Specia. 2010. Translating from complex to simplified sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language (PROPOR'10)*.

Sara Stymne, Jörg Tiedemann, Christian Hardmeier, and Joakim Nivre. 2013. Statistical machine translation with readability constraints. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*.

K. Toutanova, D. Klein, C. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL*, pages 252–259, Edmonton, Canada.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *In Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163—-173.

Sowmya Vajjala and Detmar Meurers. 2013. On the applicability of readability models to web texts. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*.

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)*.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of ACL 2012*.

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*.

Sanja Štajner, Biljana Drndarevic, and Horaccio Saggion. 2013. Corpus-based sentence deletion and split decisions for spanish text simplification. In *CICLing 2013: The 14th International Conference on Intelligent Text Processing and Computational Linguistics*.