Identifying Targets for Syntactic Simplification

Julie Medero, Mari Ostendorf

Department of Electrical Engineering, University of Washington, Seattle

{jmedero,ostendor}@u.washington.edu

Abstract

Most work on automatic text simplification considers lexical difficulty separate from syntactic simplification. In this study, we use both factors together to predict a variety of sentence changes, including the standard problems of splitting and shortening, as well as expanding to define difficult words that are important to the topic. We leverage a variety of lexical and parse features, as well as a score of the relatedness of a sentence to the topic of its document.

Index Terms: text simplification, parse complexity

1. Introduction

A number of audiences can be better served by having access to easy-to-read English texts. Young readers, second language learners, and low-literacy adults all have difficulty understanding some texts written for a general English-speaking audience. Access to simplified textbooks and other learning materials is frequently cited as crucial to the success of secondary school students with learning disabilities or very low performance [1, 2]. Creating those simplifications manually is a difficult process that teachers rarely have the time or resources to provide for their students.

One strategy for obtaining simple texts on a particular topic is to filter a set of topically similar texts by reading level. Projects like REAP [3] and Read-X [4] combine topic searches on the web with readability filters to identify level-appropriate texts for individual readers. A significant research effort has gone into automatically detecting the reading level of general documents. Traditional readability measures like the Flesch-Kincade Grade Level index [5] and the Gunning Fog index [6] rely on easily-calculated approximations to complexity based on features like sentence length and syllable counts. The Coh-Metrix project includes features based on words, POS tags, argument overlap and topic cohesion, which it measures through sentence and paragraph similarity measures using LSA [7]. Newer systems apply more advanced language modeling and statistical learning to the task [8, 9, 3]. While these filtering approaches have seen success, there may be many instances where texts at the appropriate level do not exist for a topic. Further, in a classroom setting with students with varying reading abilities, a teacher may want to simplify the same text to different levels, providing parallel texts in terms of content to all of their students to enhance discussions and full-class activities.

The recent availability of electronic texts written for a lower reading level in Simple Wikipedia[1] provides data from which researchers can learn conditions for automatic simplification. One study [10] uses the occurrence rate of words in simple vs. standard Wikipedia articles as an indicator of word difficulty and leverages features of Wiktionary definitions to predict word-level (lexical) targets of simplification. Of course, this definition of difficulty does not account for the fact that some difficult words are in the simple articles when important to the topic being discussed, e.g.medical terms. Another study leverages the simple and standard Wikipedia article to classify whole sentences as "simple" or "original" with labels according to which source it came from [11], acknowledging the problem that standard Wikipedia articles can include simple sentences. They use both lexical and parse features but without distinguishing between types of simplification targets (e.g. lexical vs. syntactic difficulty). They find that training models for specific categories of articles gives better performance, which lends some support our hypothesis that topicality of words plays a role in the simplification strategy. Yatskar and colleagues [12] use Wikipedia edit logs to extract pairs of simple and difficult words and phrases, using editor comments to more reliably identify edits aimed at simplification for training. Human judges are used in evaluation.

Other work on sentence simplification has focused on making sentences easier for machines to process in downstream applications. In multi-document summarization, removing certain types of syntactic constructions (e.g. appositives, gerundive clauses, non-restrictive relative clauses) can shorten sentences, which helps to keep extractive summaries within proscribed length limits [13, 14]. Tasks like semantic role labeling [15] and automatic question generation [16] can also be aided by defining local transformations of syntactic trees. A crucial difference between the goals of simplification for computer language processing vs. human readers is that for human readers the handling of difficult words is an important issue, not just syntactic complexity. In such cases, longer texts may be preferable if they provide an explanation for a difficult word.

This work considers the task of identifying sentences that should be targeted for syntactic simplification, building on initial analyses of a sentence-aligned corpus of original and simplified news articles [17]. A key difference from other work is the prediction of cases where sentences are expanded, as well compression methods of simplification. The rest of the paper is organizes as follows. Section 2 describes the classification task and data used in this study. Section 3 describes the features and experimental setup. Section 4 summarizes the results, and Section 5 concludes with a discussion of the results and future work.

2. Classification Task

Human editors may make any of a large number of syntactic changes to a text to improve its readability. Here, we examine three types of changes: splits, omissions, and expansions.

The **Split** class includes all instances in which one original sentence is split into two or more simplified sentences, e.g.:

(ROOT (S In the face of deregulation utilities here quit

[1] http://simple.wikipedia.org

building power plants, (S limiting supply), while (S demand kept (S going up)).))

SPLIT $\longrightarrow$

(ROOT (S (S Utilities quit (S building power plants)), and so (S the supply was limited).))
(ROOT (S The demand kept (S going up).))

The **Omit** class includes cases in which the simplified sentence has fewer S nodes than the original. It generally means that some content has been dropped from the original sentence. In the following example, the attribution is removed:

(ROOT (S Jane Garvey says, (S It's important (S to note that (S aviation is growing))).))

OMIT $\longrightarrow$

(ROOT (S Also, more planes are flying.))

The reverse of an Omit is an **Expand**, which occurs when the total number of S nodes increases. Expanded sentences tend to be ones where additional context or explanation is added, or were complex syntactic structures are replaced by conjoined S nodes. An example of the former case:

(ROOT (S It's becoming a form of mass transportation for a number of people.))

EXPAND $\longrightarrow$

(ROOT (S Airplanes are a form of mass transportation now, like (S trains and buses were in the past.)))

These changes are defined in terms of the number of sentences and S nodes within a sentence given hand alignments at the sentence level. The definitions assume that all differences in S node count are a direct result of simplification. In fact, there are some cases where the comparable (but not parallel) texts we are working with simply include slightly different information. Manual inspection suggests that this is relatively rare, and we neglect this "noise" in the labels.

3. Experiments

3.1. Methods

We use 1988 aligned sentence pairs from the dataset used in [17], including articles from the full and elementary versions of Encyclopedia Britannica from [18], and full and abridged versions of CNN news articles from the Western/Pacific Literacy Network website.[2] All sentences are automatically labeled using the definitions described above and the output of an automatic sentence segmenter [19] and syntactic parser [20]. Instances where multiple original sentences map to a single simplified sentence are excluded for the current analysis since they are rare (making up less than 3% of our total data) and are not compatible with our sentence-level classifier, but they should eventually be folded into any analysis of content omissions as part of the simplification process. The number of sentences of each type is given in Table 1.

We use icsiboost [21] to perform 10-fold cross-validation and anti-prior weighting. Two classifiers are designed to label sentences: a binary Split vs. No Split classifier, and a 3-way Expand vs. Omit vs. No Change Classifier. To examine the influence of different individual features, we hold out one partition for dev, one for tuning, and one for testing for each fold. Feature pruning is done by dropping features that are not chosen in at least 4 out of 10 training folds, which reduces our feature set size from 234 to 44. ROC curves are based on test set results.

[2]http://literacynet.org/cnnsf/, accessed June 15, 2004

	Omit	Expand	No Change
Split	126	264	107
No Split	478	192	821

Table 1: Distribution of labels for all sentences.

3.2. Features

The predictions in this work represent types of sentence transformations, so features extracted are at the sentence level. We use multiple lexical and parse-based feature statistics to capture word difficulty vs. syntactic complexity, respectively, and a topicality score to help with distinguishing omit vs. expand cases. The feature extraction modules leverage several online resources, as described below.

Lexical Features. Identifying difficult words is important in simplification, as evidenced by the dominant role of lexical features in reading level detection. Unigram frequency is well known to correlate with reading difficulty; here, we estimate unigrams based on the Corpus of Contemporary American English (COCA) [22]. To compute a sentence-level statistic, we count the number of word tokens in the sentence with COCA frequency of less than some threshold f_t, considering thresholds corresponding to several different points in the cumulative distribution of words in COCA ranging over [0.5,.35]. In addition to unigram frequencies, we add word-level features from Wiktionary,[3] which we have previously found promising for identifying difficult words [10]. The total number of senses and total number of translations is extracted for each word, and then those counts are binned in the same way as the unigram frequencies. Finally, the number of words that are and are not on the Simple English Wikipedia vocabulary list from the Basic English Institute (BEI)[4] are included.

Syntactic Features We expect syntactic features to be important for identifying constructions that are likely to be split or omitted. Syntactic features are extracted from the syntactic parse trees of the original (unsimplified) sentences, as generated by the Stanford parser [20]. Four types of syntactic features are used: i) constituent counts, ii) dependent pair counts, and iii) ROOT $\rightarrow$preterminal n-gram scores.

It is expected that sentences with embedded S and $SBAR$ nodes may be more likely to be split, or to have pieces omitted. For each sentence, the number of S and $SBAR$ nodes are counted, along with the total number CC nodes, the total number of words and the total number of nonterminal nodes.

Expecting that unusual syntactic constructions will tend to be simplified, we characterize unusual constructions in two ways, leveraging statistics from the English Wikipedia.[5] First, we generate counts of dependency pairs of nonterminal nodes from syntactic parses, and as with unigrams, we calculate frequency thresholds based on cumulative distributions in Wikipedia. For each sentence, we include counts of dependency pairs with frequencies less than each frequency threshold. In addition, we build a trigram language model based on non-terminal paths from the ROOT node to each pre-terminal node in the Wikipedia corpus. For each sentence, we include the perplexity of the set of ROOT$\rightarrow$pre-terminal paths for that sentence's parse as a feature. Finally, we include the number of productions in the syntactic parse with a branching factor of at

[3]http://en.wiktionary.org
[4]http://www.basic-english.org/
[5]http://en.wikipedia.org

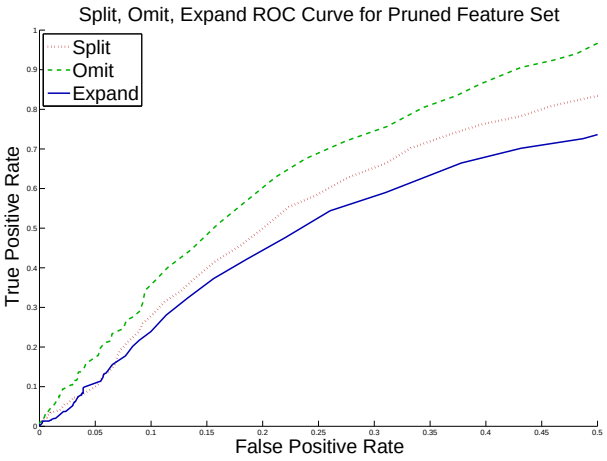

Figure 1: ROC curve for predicting splits, omits, and expands

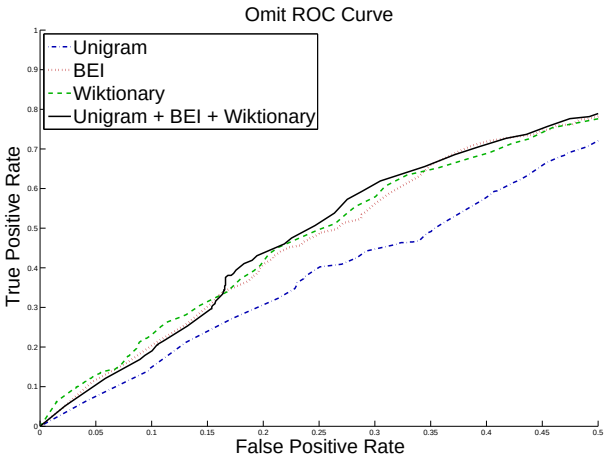

Figure 2: Performance on Omit classification for different sets of lexical features

Omit	Expand	Split
Constituent Count	Constituent Count	Parse Paths
Topic Score	Topic Score	Topic Score
Parse Paths	Position in Document	Wiktionary

Table 2: Top features for predicting Splits, Omits and Expands

least n, where $n = 1, \ldots, 7$.

Topicality Score. The relatedness of words to the topic of an article is important for deciding whether they are sufficiently central to keep, in which case they may require expansion. Less central difficult words are more likely to be excluded or paraphrased during simplification. We calculate sentence-to-document similarity scores analogous to the textual cohesion scores used in Coh-metrix [7], but using a pLSA space instead of LSA. The pLSA representation of each sentence and article is a vector of posterior probabilities from 50 unigram language models from subsets of the New York Times section of the Gigaword corpus [23]. The partitions of the dataset come from k-means clustering initialized by term-document clustering output by CLuTO [24]. We calculate the cosine distance between each sentence and 1) the previous sentence; 2) all previous sentences in the document; and 3) the full article.

4. Results

4.1. Overall Performance

Overall, the prediction of expansions is more difficult than splits, and we perform best on predicting omissions. ROC curves for each classifier using the pruned feature set are shown in Figure 1. At the equal error rate point, performance ranges from 0.27 for omit to 0.36 for expansions. Performance using the full set of 234 features was comparable to or slightly worse than the pruned feature set. The most important features for detecting the simplification phenomena are given in Table 2.

4.2. Feature Analysis

Lexical Features. We explore the extent to which word-based information from sources like the BEI word list and Wiktionary are useful in comparison to and in combination with unigram frequencies. Figure 2 shows that both the BEI word list features and the Wiktionary-based word features outperform unigram features for predicting omissions. At a false oisutuve rate of 0.25, differences are significant with $p < 0.0001$. Combining all three feature sets slightly outperforms any single feature set at high false positive rates, but the differences are small. Results were similar but smaller for labeling Splits, and lexical features were not successful in general in predicting Expands.

Syntactic Features. Unsurprisingly, the number of S nodes was the most-used syntactic feature in predicting all three changes, along with the total sentence length in words. The total number of words and the number of nodes with a branching factor greater than 3 was also commonly used. Number of $SBAR$ and CC nodes were not commonly used predictors.

We expected that rare syntactic structures, as captured by dependency pairs and our trigram language model, would be good predictors of simplification. These features were rarely used by the classifier, though. The perplexity of the syntactic language model was used more frequently than the dependency pair-based features.

Topicality Score. To examine the potential usefulness of our pLSA-based topicality features, we look at the distribution the different types of sentence changes as a function of the topic score quantifying relatedness of the candidate sentence to the document, shown in Figure 3. As we hypothesized, sentences with a high topic score (more centrally related to the document) that are transformed are more likely to have expansions than those with low scores, and sentences with a low topic score are more likely to be omitted.

5. Discussion and Future Work

We have provided a description of an initial system aimed at identifying and describing syntactic changes made during manual text simplification. While the features we present, including topicality features and expanded lexical features, seem promising, there is still room for improvement in characterizing the conditions in which splits and expansions take place, including looking at features from more general work on readability [25].

Previous work and our own analysis indicated that three common changes in simplification are replacing or explaining difficult words and phrases, removing extraneous details, and separating syntactically complex sentences into multiple shorter sentences. Here, we focus on those changes, as identified by

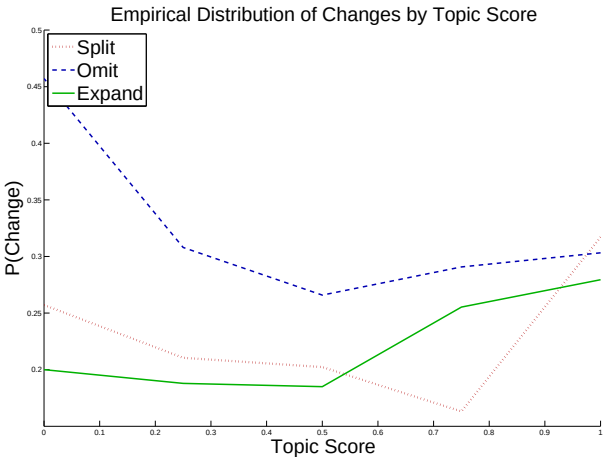

Figure 3: Relative frequency of sentence changes as function of topic score.

changes in the count of S nodes, because they are easy to automate given sentence aligned data. Future work will explore ways to characterize and learn other types of changes that human simplifiers make, including converting passive voice to active voice, removing attribution ("John Smith said that..."), and replacing pronouns with their antecedents.

6. Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-0718124 and by the National Science Foundation under Grant No. IIS-0916951. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation

7. References

[1] T. C. Lovitt and S. V. Horton, "Strategies for adapting science textbooks for youth with learning disabilities," *Remedial and Special Education*, vol. 15, pp. 105–116, 1994.

[2] J. S. Schumm and K. Stickler, "Guidelines for adapting content area textbooks: Keeping teachers and students content," *Intervention in School and Clinic*, 1991.

[3] M. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi, "Classroom success of an intelligent tutoring system for lexical practice and reading comprehension," in *Proc. ICSLP*, 2006.

[4] E. Miltsakaki and A. Troutt, "Real-time web text classification and analysis of reading difficulty," in *Third Workshop on Innovative Use of NLP for Building Educational Applications*, 2008.

[5] J. J. P. Kincaid, R. Fishburne, R. Rodgers, and B. Chisson, "Derivation of new readability formulas for Navy enlisted personnel," *Research Branch Report 8-75*, 1975.

[6] R. Grunning, *The Technique of Clear Writing*. McGraw-Hill, 1952.

[7] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai, "Coh-metrix:analysis of text on cohesion and language," *Behavior Research Methods, Instruments, and Computers*, vol. 36, pp. 193–202, 2004.

[8] S. K. S. Sharoff and A. Hartley, "Seeking needles in the web haystack: Finding texts suitable for language learners," in *TaLC-8*, 2008.

[9] S. Petersen and M. Ostendorf, "A machine learning approach to reading level assessment," *Computer, Speech and Language*, 2009.

[10] J. Medero and M. Ostendorf, "Analysis of vocabulary difficulty using wiktionary," in *Proc. SLaTE Workshop*, 2009.

[11] C. Napoles and M. Dredze, "Learning simple wikipedia: a cogitation in ascertaining abecedarian language," in *Proc. NAACL HLT Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, 2010, pp. 42–50.

[12] M. Yatskar, B. Pang, C. D. N. Mizil, and L. Lee, "For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia," in *Proc. NAACL HLT Conference*, Los Angeles, California, 2010, pp. 365–368.

[13] L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova, "Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion," *Inf. Process. Manage.*, vol. 43, pp. 1606–1618, November 2007.

[14] A. Siddharthan, A. Nenkova, and K. McKeown, "Syntactic simplification for improving content selection in multi-document summarization," in *Proc. ACL*, 2004, pp. 407–414.

[15] D. Vickrey and D. Koller, "Sentence simplification for semantic role labeling," in *Proc. ACL-HLT*, 2008, pp. 344–352.

[16] M. Heilman and N. A. Smith, "Tree edit models for recognizing textal entailments, paraphrases and answers to questions," in *Proc. NAACL-HLT*, 2010.

[17] S. E. Petersen and M. Ostendorf, "Text simplification for language learners: A corpus analysis," in *Proc. SLaTE Workshop*, 2007, pp. 69–72.

[18] R. Barzilay and N. Elhadad, "Sentence alignment for monolingual comparable corpora," in *Proc. EMNLP*, 2003, p. 2532.

[19] A. Ratnaparkhi, "A maximum entropy part-of-speech tagger," in *Proc. Empirical Methods in Natural Language Processing Conference*, 1996, pp. 133–141.

[20] D. Klein and C. Manning, "Accurate unlexicalized parsing," in *Proc. ACL*, 2003, pp. 423–430.

[21] B. Favre, D. Hakkani-Tür, and S. Cuendet, "Icsiboost," http://code.google.come/p/icsiboost, 2007.

[22] M. Davies, "The Corpus of Contemporary American English (COCA): 425 million words, 1990-present," http://www.americancorpus.org, 2008-.

[23] D. Graff and C. Cieri, "English gigaword," 2003.

[24] Y. Zhao and G.Karypis, "Hierarchical clustering algorithms for document datasets," *Data Mining and Knowledge Discovery*, vol. 10, pp. 141–168, 2005.

[25] E. Pitler and A. Nenkova, "Revisiting readability: A unified framework for predicting text quality," in *Proc. EMNLP*, 2008.