Simple English Wikipedia: A New Text Simplification Task

William Coster
Computer Science Department
Pomona College
Claremont, CA 91711
wpc02009@pomona.edu

David Kauchak
Computer Science Department
Pomona College
Claremont, CA 91711
dkauchak@cs.pomona.edu

Abstract

In this paper we examine the task of sentence simplification which aims to reduce the reading complexity of a sentence by incorporating more accessible vocabulary and sentence structure. We introduce a new data set that pairs English Wikipedia with Simple English Wikipedia and is orders of magnitude larger than any previously examined for sentence simplification. The data contains the full range of simplification operations including rewording, reordering, insertion and deletion. We provide an analysis of this corpus as well as preliminary results using a phrase-based translation approach for simplification.

1 Introduction

The task of text simplification aims to reduce the complexity of text while maintaining the content (Chandrasekar and Srinivas, 1997; Carroll et al., 1998; Feng, 2008). In this paper, we explore the *sentence* simplification problem: given a sentence, the goal is to produce an equivalent sentence where the vocabulary and sentence structure are simpler.

Text simplification has a number of important applications. Simplification techniques can be used to make text resources available to a broader range of readers, including children, language learners, the elderly, the hearing impaired and people with aphasia or cognitive disabilities (Carroll et al., 1998; Feng, 2008). As a preprocessing step, simplification can improve the performance of NLP tasks, including parsing, semantic role labeling, machine translation and summarization (Miwa et al., 2010; Jonnalagadda et al., 2009; Vickrey and Koller, 2008; Chandrasekar and Srinivas, 1997). Finally, models for text simplification are similar to models for sentence compression; advances in simplification can benefit compression, which has applications in mobile devices, summarization and captioning (Knight and Marcu, 2002; McDonald, 2006; Galley and McKeown, 2007; Nomoto, 2009; Cohn and Lapata, 2009).

One of the key challenges for text simplification is data availability. The small amount of simplification data currently available has prevented the application of data-driven techniques like those used in other text-to-text translation areas (Och and Ney, 2004; Chiang, 2010). Most prior techniques for text simplification have involved either hand-crafted rules (Vickrey and Koller, 2008; Feng, 2008) or learned within a very restricted rule space (Chandrasekar and Srinivas, 1997).

We have generated a data set consisting of 137K aligned simplified/unsimplified sentence pairs by pairing documents, then sentences from English Wikipedia[1] with corresponding documents and sentences from Simple English Wikipedia[2]. Simple English Wikipedia contains articles aimed at children and English language learners and contains similar content to English Wikipedia but with simpler vocabulary and grammar.

Figure 1 shows example sentence simplifications from the data set. Like machine translation and other text-to-text domains, text simplification involves the full range of transformation operations including deletion, rewording, reordering and insertion.

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

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

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 665–669,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

a.	Normal:	As Isolde arrives at his side, Tristan dies *with her name on his lips*.
	Simple:	As Isolde arrives at his side, Tristan dies *while speaking her name*.
b.	Normal:	Alfonso Perez *Munoz, usually referred to as Alfonso,* is a former Spanish *footballer, in the striker position*.
	Simple:	Alfonso Perez is a former Spanish *football player*.
c.	Normal:	Endemic types *or species* are *especially* likely to develop on islands because *of their geographical isolation*.
	Simple:	Endemic types are *most* likely to develop on islands because *they are isolated*.
d.	Normal:	The reverse process, producing electrical energy from mechanical, energy, is accomplished by a generator or dynamo.
	Simple:	A dynamo or an electric generator does the reverse: it changes mechanical movement into electric energy.

Figure 1: Example sentence simplifications extracted from Wikipedia. *Normal* refers to a sentence in an English Wikipedia article and *Simple* to a corresponding sentence in Simple English Wikipedia.

2 Previous Data

Wikipedia and Simple English Wikipedia have both received some recent attention as a useful resource for text simplification and the related task of text compression. Yamangil and Nelken (2008) examine the history logs of English Wikipedia to learn sentence compression rules. Yatskar et al. (2010) learn a set of candidate phrase simplification rules based on edits identified in the revision histories of both Simple English Wikipedia and English Wikipedia. However, they only provide a list of the top phrasal simplifications and do not utilize them in an end-to-end simplification system. Finally, Napoles and Dredze (2010) provide an analysis of the differences between documents in English Wikipedia and Simple English Wikipedia, though they do not view the data set as a parallel corpus.

Although the simplification problem shares some characteristics with the text compression problem, existing text compression data sets are small and contain a restricted set of possible transformations (often only deletion). Knight and Marcu (2002) introduced the Zipf-Davis corpus which contains 1K sentence pairs. Cohn and Lapata (2009) manually generated two parallel corpora from news stories totaling 3K sentence pairs. Finally, Nomoto (2009) generated a data set based on RSS feeds containing 2K sentence pairs.

3 Simplification Corpus Generation

We generated a parallel simplification corpus by aligning sentences between English Wikipedia and Simple English Wikipedia. We obtained complete copies of English Wikipedia and Simple English Wikipedia in May 2010. We first paired the articles by title, then removed all article pairs where either article: contained only a single line, was flagged as a stub, was flagged as a disambiguation page or was a meta-page about Wikipedia. After pairing and filtering, 10,588 aligned, content article pairs remained (a 90% reduction from the original 110K Simple English Wikipedia articles). Throughout the rest of this paper we will refer to unsimplified text from English Wikipedia as *normal* and to the simplified version from Simple English Wikipedia as *simple*.

To generate aligned sentence pairs from the aligned document pairs we followed an approach similar to those utilized in previous monolingual alignment problems (Barzilay and Elhadad, 2003; Nelken and Shieber, 2006). Paragraphs were identified based on formatting information available in the articles. Each simple paragraph was then aligned to every normal paragraph where the TF-IDF, cosine similarity was over a threshold or 0.5. We initially investigated the paragraph clustering preprocessing step in (Barzilay and Elhadad, 2003), but did not find a qualitative difference and opted for the simpler similarity-based alignment approach, which does not require manual annotation.

For each aligned paragraph pair (i.e. a simple paragraph and one or more normal paragraphs), we then used a dynamic programming approach to find that best global sentence alignment following Barzilay and Elhadad (2003). Specifically, given n normal sentences to align to m simple sentences, we find $a(n, m)$ using the following recurrence:

$$a(i,j) = \max \begin{cases} a(i,j-1) - skip\_penalty \\ a(i-1,j) - skip\_penalty \\ a(i-1,j-1) + sim(i,j) \\ a(i-1,j-2) + sim(i,j) + sim(i,j-1) \\ a(i-2,j-1) + sim(i,j) + sim(i-1,j) \\ a(i-2,j-2) + sim(i,j-1) + sim(i-1,j) \end{cases}$$

where each line above corresponds to a sentence alignment operation: skip the simple sentence, skip the normal sentence, align one normal to one simple, align one normal to two simple, align two normal to one simple and align two normal to two simple. $sim(i, j)$ is the similarity between the ith normal sentence and the jth simple sentence and was calculated using TF-IDF, cosine similarity. We set $skip\_penalty = 0.0001$ manually.

Barzilay and Elhadad (2003) further discourage aligning dissimilar sentences by including a "mismatch penalty" in the similarity measure. Instead, we included a filtering step removing all sentence pairs with a normalized similarity below a threshold of 0.5. We found this approach to be more intuitive and allowed us to compare the effects of differing levels of similarity in the training set. Our choice of threshold is high enough to ensure that most alignments are correct, but low enough to allow for variation in the paired sentences. In the future, we hope to explore other similarity techniques that will pair sentences with even larger variation.

4 Corpus Analysis

From the 10K article pairs, we extracted 75K aligned paragraphs. From these, we extracted the final set of 137K aligned sentence pairs. To evaluate the quality of the aligned sentences, we asked two human evaluators to independently judge whether or not the aligned sentences were correctly aligned on a random sample of 100 sentence pairs. They then were asked to reach a consensus about correctness. 91/100 were identified as correct, though many of the remaining 9 also had some partial content overlap. We also repeated the experiment using only those sentences with a similarity above 0.75 (rather than 0.50 in the original data). This reduced the number of pairs from 137K to 90K, but the evaluators identified 98/100 as correct. The analysis throughout the rest of the section is for threshold of 0.5, though similar results were also seen for the threshold of 0.75.

Although the average simple article contained approximately 40 sentences, we extracted an average of 14 aligned sentence pairs per article. Qualitatively, it is rare to find a simple article that is a *direct translation* of the normal article, that is, a simple article that was generated by only making sentence-level changes to the normal document. However, there is a strong relationship between the two data sets: 27% of our aligned sentences were identical between simple and normal. We left these identical sentence pairs in our data set since not all sentences need to be simplified and it is important for any simplification algorithm to be able to handle this case.

Much of the content without direct correspondence is removed during paragraph alignment. 65% of the simple paragraphs do not align to a normal paragraphs and are ignored. On top of this, within aligned paragraphs, there are a large number of sentences that do not align. Table 1 shows the proportion of the different sentence level alignment operations in our data set. On both the simple and normal sides there are many sentences that do not align.

Operation	%
skip simple	27%
skip normal	23%
one normal to one simple	37%
one normal to two simple	8%
two normal to one simple	5%

Table 1: Frequency of sentence-level alignment operations based on our learned sentence alignment. No 2-to-2 alignments were found in the data.

To better understand how sentences are transformed from normal to simple sentences we learned a word alignment using GIZA++ (Och and Ney, 2003). Based on this word alignment, we calculated the percentage of sentences that included: **re-**

wordings – a normal word is changed to a different simple word, **deletions** – a normal word is deleted, **reorderings** – non-monotonic alignment, **splits** – a normal words is split into multiple simple words, and **merges** – multiple normal words are condensed to a single simple word.

Transformation	%
rewordings	65%
deletions	47%
reorders	34%
merges	31%
splits	27%

Table 2: Percentage of sentence pairs that contained word-level operations based on the induced word alignment. Splits and merges are from the perspective of words in the normal sentence. These are not mutually exclusive events.

Table 2 shows the percentage of each of these phenomena occurring in the sentence pairs. All of the different operations occur frequently in the data set with rewordings being particularly prevalent.

5 Sentence-level Text Simplification

To understand the usefulness of this data we ran preliminary experiments to learn a sentence-level simplification system. We view the problem of text simplification as an English-to-English translation problem. Motivated by the importance of lexical changes, we used Moses, a phrase-based machine translation system (Och and Ney, 2004).[3] We trained Moses on 124K pairs from the data set and the n-gram language model on the simple side of this data. We trained the hyper-parameters of the log-linear model on a 500 sentence pair development set.

We compared the trained system to a baseline of not doing any simplification (NONE). We evaluated the two approaches on a test set of 1300 sentence pairs. Since there is currently no standard for automatically evaluating sentence simplification, we used three different automatic measures that have been used in related domains: BLEU, which has been used extensively in machine translation (Papineni et al., 2002), and word-level F1 and simple string accuracy (SSA) which have been suggested for text compression (Clarke and Lapata, 2006). All three of these measures have been shown to correlate with human judgements in their respective domains.

System	BLEU	word-F1	SSA
NONE	0.5937	0.5967	0.6179
Moses	0.5987	0.6076	0.6224
Moses-Oracle	0.6317	0.6661	0.6550

Table 3: Test scores for the baseline (NONE), Moses and Moses-Oracle.

Table 3 shows the results of our initial test. All differences are statistically significant at $p = 0.01$, measured using bootstrap resampling with 100 samples (Koehn, 2004). Although the baseline does well (recall that over a quarter of the sentence pairs in the data set are identical) the phrase-based approach does obtain a statistically significant improvement.

To understand the the limits of the phrase-based model for text simplification, we generated an n-best list of the 1000 most-likely simplifications for each test sentence. We then greedily picked the simplification from this n-best list that had the highest sentence-level BLEU score based on the test examples, labeled Moses-Oracle in Table 3. The large difference between Moses and Moses-Oracle indicates possible room for improvement utilizing better parameter estimation or n-best list reranking techniques (Och et al., 2004; Ge and Mooney, 2006).

6 Conclusion

We have described a new text simplification data set generated from aligning sentences in Simple English Wikipedia with sentences in English Wikipedia. The data set is orders of magnitude larger than any currently available for text simplification or for the related field of text compression and is publicly available.[4] We provided preliminary text simplification results using Moses, a phrase-based translation system, and saw a statistically significant improvement of 0.005 BLEU over the baseline of no simplification and showed that further improvement of up to 0.034 BLEU may be possible based on the oracle results. In the future, we hope to explore alignment techniques more tailored to simplification as well as applications of this data to text simplification.

[3] We also experimented with T3 (Cohn and Lapata, 2009) but the results were poor and are not presented here.

[4] http://www.cs.pomona.edu/∼dkauchak/simplification/

References

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of EMNLP*.

John Carroll, Gido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of AAAI Workshop on Integrating AI and Assistive Technology*.

Raman Chandrasekar and Bangalore Srinivas. 1997. Automatic induction of rules for text simplification. In *Knowledge Based Systems*.

David Chiang. 2010. Learning to translate with source and target syntax. In *Proceedings of ACL*.

James Clarke and Mirella Lapata. 2006. Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *Proceedings of ACL*.

Trevor Cohn and Mirella Lapata. 2009. Sentence compression as tree transduction. *Journal of Artificial Intelligence Research*.

Lijun Feng. 2008. Text simplification: A survey. CUNY Technical Report.

Michel Galley and Kathleen McKeown. 2007. Lexicalized Markov grammars for sentence compression. In *Proceedings of HLT/NAACL*.

Ruifang Ge and Raymond Mooney. 2006. Discriminative reranking for semantic parsing. In *Proceedings of COLING*.

Siddhartha Jonnalagadda, Luis Tari, Jorg Hakenberg, Chitta Baral, and Graciela Gonzalez. 2009. Towards effective sentence simplification for automatic processing of biomedical text. In *Proceedings of HLT/NAACL*.

Dan Klein and Christopher Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL*.

Kevin Knight and Daniel Marcu. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*.

Ryan McDonald. 2006. Discriminative sentence compression with soft syntactic evidence. In *Proceedings of EACL*.

Makoto Miwa, Rune Saetre, Yusuke Miyao, and Jun'ichi Tsujii. 2010. Entity-focused sentence simplication for relation extraction. In *Proceedings of COLING*.

Courtney Napoles and Mark Dredze. 2010. Learning simple Wikipedia: A cogitation in ascertaining abecedarian language. In *Proceedings of HLT/NAACL Workshop on Computation Linguistics and Writing*.

Rani Nelken and Stuart Shieber. 2006. Towards robust context-sensitive sentence alignment for monolingual corpora. In *Proceedings of AMTA*.

Tadashi Nomoto. 2007. Discriminative sentence compression with conditional random fields. In *Information Processing and Management*.

Tadashi Nomoto. 2008. A generic sentence trimmer with CRFs. In *Proceedings of HLT/NAACL*.

Tadashi Nomoto. 2009. A comparison of model free versus model intensive approaches to sentence compression. In *Proceedings of EMNLP*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Franz Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*.

Franz Josef Och, Kenji Yamada, Stanford U, Alex Fraser, Daniel Gildea, and Viren Jain. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of HLT/NAACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Emily Pitler. 2010. Methods for sentence compression. Technical Report MS-CIS-10-20, University of Pennsylvania.

Jenine Turner and Eugene Charniak. 2005. Supervised and unsupervised learning for sentence compression. In *Proceedings of ACL*.

David Vickrey and Daphne Koller. 2008. Sentence simplification for semantic role labeling. In *Proceedings of ACL*.

Elif Yamangil and Rani Nelken. 2008. Mining Wikipedia revision histories for improving sentence compression. In *ACL*.

Mark Yatskar, Bo Pang, Critian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *HLT/NAACL Short Papers*.