Sentence Simplification by Monolingual Machine Translation

Sander Wubben
Tilburg University
P.O. Box 90135
5000 LE Tilburg
The Netherlands
s.wubben@uvt.nl

Antal van den Bosch
Radboud University Nijmegen
P.O. Box 9103
6500 HD Nijmegen
The Netherlands
a.vandenbosch@let.ru.nl

Emiel Krahmer
Tilburg University
P.O. Box 90135
5000 LE Tilburg
The Netherlands
e.j.krahmer@uvt.nl

Abstract

In this paper we describe a method for simplifying sentences using Phrase Based Machine Translation, augmented with a re-ranking heuristic based on dissimilarity, and trained on a monolingual parallel corpus. We compare our system to a word-substitution baseline and two state-of-the-art systems, all trained and tested on paired sentences from the English part of Wikipedia and Simple Wikipedia. Human test subjects judge the output of the different systems. Analysing the judgements shows that by relatively careful phrase-based paraphrasing our model achieves similar simplification results to state-of-the-art systems, while generating better formed output. We also argue that text readability metrics such as the Flesch-Kincaid grade level should be used with caution when evaluating the output of simplification systems.

1 Introduction

Sentence simplification can be defined as the process of producing a simplified version of a sentence by changing some of the lexical material and grammatical structure of that sentence, while still preserving the semantic content of the original sentence, in order to ease its understanding. Particularly language learners (Siddharthan, 2002), people with reading disabilities (Inui et al., 2003) such as aphasia (Carroll et al., 1999), and low-literacy readers (Watanabe et al., 2009) can benefit from this application. It can serve to generate output in a specific limited format, such as subtitles (Daelemans et al., 2004). Sentence simplification can also serve to preprocess the input of other tasks, such as summarization (Knight and Marcu, 2000), parsing, machine translation (Chandrasekar et al., 1996), semantic role labeling (Vickrey and Koller, 2008) or sentence fusion (Filippova and Strube, 2008).

The goal of simplification is to achieve an improvement in readability, defined as the ease with which a text can be understood. Some of the factors that are known to help increase the readability of text are the vocabulary used, the length of the sentences, the syntactic structures present in the text, and the usage of discourse markers. One effort to create a simple version of English at the vocabulary level has been the creation of Basic English by Charles Kay Ogden. Basic English is a controlled language with a basic vocabulary consisting of 850 words. According to Ogden, 90 percent of all dictionary entries can be paraphrased using these 850 words. An example of a resource that is written using mainly Basic English is the English Simple Wikipedia. Articles on English Simple Wikipedia are similar to articles found in the traditional English Wikipedia, but written using a limited vocabulary (using Basic English where possible). Generally the structure of the sentences in English Simple Wikipedia is less complicated and the sentences are somewhat shorter than those found in English Wikipedia; we offer more detailed statistics below.

1.1 Related work

Most earlier work on sentence simplification adopted rule-based approaches. A frequently applied type of rule, aimed to reduce overall sentence length, splits long sentences on the basis of syntactic

information (Chandrasekar and Srinivas, 1997; Carroll et al., 1998; Canning et al., 2000; Vickrey and Koller, 2008). There has also been work on lexical substitution for simplification, where the aim is to substitute difficult words with simpler synonyms, derived from WordNet or dictionaries (Inui et al., 2003).

Zhu et al. (2010) examine the use of paired documents in English Wikipedia and Simple Wikipedia for a data-driven approach to the sentence simplification task. They propose a probabilistic, syntax-based machine translation approach to the problem and compare against a baseline of no simplification and a phrase-based machine translation approach. In a similar vein, Coster and Kauchak (2011) use a parallel corpus of paired documents from Simple Wikipedia and Wikipedia to train a phrase-based machine translation model coupled with a deletion model. Another useful resource is the edit history of Simple Wikipedia, from which simplifications can be learned (Yatskar et al., 2010). Woodsend and Lapata (2011) investigate the use of Simple Wikipedia edit histories and an aligned Wikipedia–Simple Wikipedia corpus to induce a model based on quasi-synchronous grammar. They select the most appropriate simplification by using integer linear programming.

We follow Zhu et al. (2010) and Coster and Kauchak (2011) in proposing that sentence simplification can be approached as a monolingual machine translation task, where the source and target languages are the same and where the output should be simpler in form from the input but similar in meaning. We differ from the approach of Zhu et al. (2010) in the sense that we do not take syntactic information into account; we rely on PBMT to do its work and implicitly learn simplifying paraphrasings of phrases. Our approach differs from Coster and Kauchak (2011) in the sense that instead of focusing on deletion in the PBMT decoding stage, we focus on dissimilarity, as simplification does not necessarily imply shortening (Woodsend and Lapata, 2011), or as the Simple Wikipedia guidelines state, "simpler does not mean short"[1]. Table 1.1 shows the average sentence length and the average word length for Wikipedia and Simple Wikipedia sentences in the PWKP dataset used in this study (Zhu et al., 2010). These numbers suggest that, although the selection criteria for sentences to be included in this dataset are biased (see Section 2.2), Simple Wikipedia sentences are about 17% shorter, while the average word length is virtually equal.

	Sent. length	Token length
Simple Wikipedia	20.87	4.89
Wikipedia	25.01	5.06

Table 1: Sentence and token length statistics for the PWKP dataset (Zhu et al., 2010).

Statistical machine translation (SMT) has already been successfully applied to the related task of paraphrasing (Quirk et al., 2004; Bannard and Callison-Burch, 2005; Madnani et al., 2007; Callison-Burch, 2008; Zhao et al., 2009; Wubben et al., 2010). SMT typically makes use of large parallel corpora to train a model on. These corpora need to be aligned at the sentence level. Large parallel corpora, such as the multilingual proceedings of the European Parliament (Europarl), are readily available for many languages. Phrase-Based Machine Translation (PBMT) is a form of SMT where the translation model aims to translate longer sequences of words ("phrases") in one go, solving part of the word ordering problem along the way that would be left to the target language model in a word-based SMT system. PMBT operates purely on statistics and no linguistic knowledge is involved in the process: the phrases that are aligned are motivated statistically, rather than linguistically. This makes PBMT adaptable to any language pair for which there is a parallel corpus available. The PBMT model makes use of a translation model, derived from the parallel corpus, and a language model, derived from a monolingual corpus in the target language. The language model is typically an n-gram model with smoothing. For any given input sentence, a search is carried out producing an n-best list of candidate translations, ranked by the decoder score, a complex scoring function including likelihood scores from the translation model, and the target language model. In principle, all of this should be transportable to a data-driven machine translation account of sentence simplification, pro-

[1] http://simple.wikipedia.org/wiki/Main_Page/Introduction

vided that a parallel corpus is available that pairs text to simplified versions of that text.

1.2 This study

In this work we aim to investigate the use of phrase-based machine translation modified with a dissimilarity component for the task of sentence simplification. While Zhu et al. (2010) have demonstrated that their approach outperforms a PBMT approach in terms of Flesch Reading Ease test scores, we are not aware of any studies that evaluate PBMT for sentence simplification with human judgements. In this study we evaluate the output of Zhu et al. (2010) (henceforth referred to as 'Zhu'), Woodsend and Lapata (2011) (henceforth referred to as 'RevILP'), our PBMT based system with dissimilarity-based re-ranking (henceforth referred to as 'PBMT-R'), a word-substitution baseline, and, as a gold standard, the original Simple Wikipedia sentences. We will first discuss the baseline, followed by the Zhu system, the RevILP system, and our PBMT-R system in Section 2. We then describe the experiment with human judges in Section 3, and its results in Section 4. We close this paper by critically discussing our results in Section 5.

2 Sentence Simplification Models

2.1 Word-Substitution Baseline

The word substitution baseline replaces words in the source sentence with (near-)synonyms that are more likely according to a language model. For each noun, adjective and verb in the sentence this model takes that word and its part-of-speech tag and retrieves from WordNet all synonyms from all synsets the word occurs in. The word is then replaced by all of its synset words, and each replacement is scored by a SRILM language model (Stolcke, 2002) with probabilities that are obtained from training on the Simple Wikipedia data. The alternative that has the highest probability according to the language model is kept. If no relevant alternative is found, the word is left unchanged. We use the Memory-Based Tagger (Daelemans et al., 1996) trained on the Brown corpus to compute the part-of-speech tags. The WordNet::QueryData[2] Perl module is used to query WordNet (Fellbaum, 1998).

2.2 Zhu et al.

Zhu et al. (2010) learn a sentence simplification model which is able to perform four rewrite operations on the parse trees of the input sentences, namely substitution, reordering, splitting, and deletion. Their model is inspired by syntax-based SMT (Yamada and Knight, 2001) and consists of a language model, a translation model and a decoder. The four mentioned simplification operations together form the translation model. Their model is trained on a corpus containing aligned sentences from English Wikipedia and English Simple Wikipedia called PWKP. The PWKP dataset consists of 108,016 pairs of aligned lines from 65,133 Wikipedia and Simple Wikipedia articles. These articles were paired by following the "interlanguage link"[3]. TF*IDF at the sentence level was used to align the sentences in the different articles (Nelken and Shieber, 2006).

Zhu et al. (2010) evaluate their system using BLEU and NIST scores, as well as various readability scores that only take into account the output sentence, such as the Flesch Reading Ease test and n-gram language model perplexity. Although their system outperforms several baselines at the level of these readability metrics, they do not achieve better when evaluated with BLEU or NIST.

2.3 RevILP

Woodsend and Lapata's (2011) model is based on quasi-synchronous grammar (Smith and Eisner, 2006). Quasi-synchronous grammar generates a loose alignment between parse trees. It operates on individual sentences annotated with syntactic information in the form of phrase structure trees. Quasi-synchronous grammar is used to generate all possible rewrite operations, after which integer linear programming is employed to select the most appropriate simplification. Their model is trained on two different datasets: one containing alignments between Wikipedia and English Simple Wikipedia (AlignILP), and one containing alignments between edits in the revision history of Simple Wikipedia (RevILP). RevILP performs best according to the

[2] http://search.cpan.org/dist/WordNet-QueryData/QueryData.pm

[3] http://en.wikipedia.org/wiki/Help:Interlanguage_links

human judgements conducted in their study. They show that it achieves better scores than Zhu et al. (2010)'s system and is not scored significantly differently from English Simple Wikipedia. In this study we compare against their best performing system, the RevILP system.

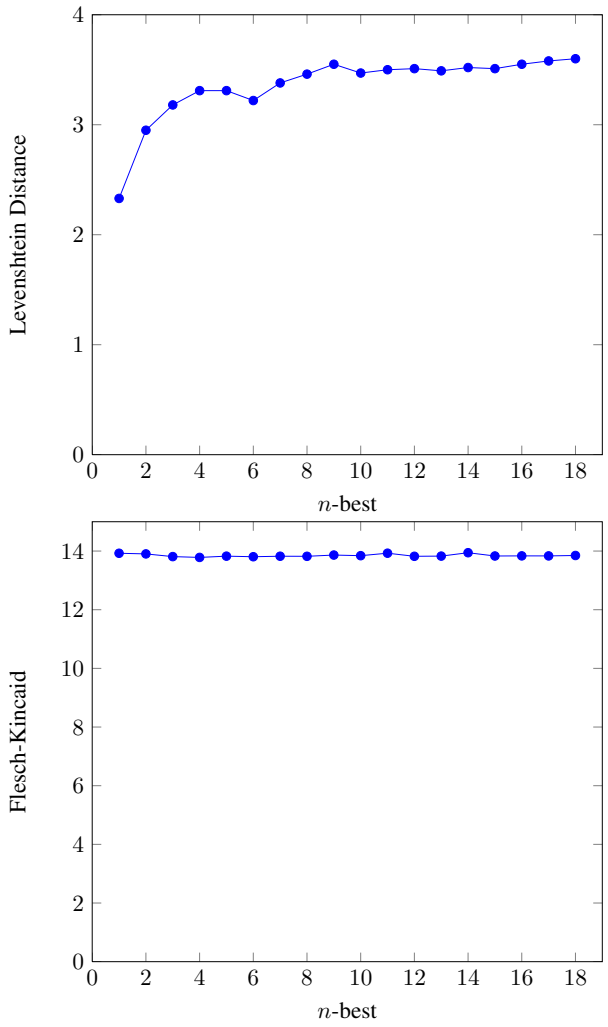

Figure 1: Levenshtein distance and Flesch-Kincaid score of output when varying the n of the n-best output of Moses.

2.4 PBMT-R

We use the Moses software to train a PBMT model (Koehn et al., 2007). The data we use is the PWKP dataset created by Zhu et al. (2010). In general, a statistical machine translation model finds a best translation $\tilde{e}$ of a text in language f to a text in language e by combining a translation model that finds the most likely translation $p(f|e)$ with a language model that outputs the most likely sentence $p(e)$:

$$\tilde{e} = arg\max_{e \in e^*} p(f|e)p(e)$$

The GIZA++ statistical alignment package is used to perform the word alignments, which are later combined into phrase alignments in the Moses pipeline (Och and Ney, 2003) to build the sentence simplification model. GIZA++ utilizes IBM Models 1 to 5 and an HMM word alignment model to find statistically motivated alignments between words. We first tokenize and lowercase all data and use all unique sentences from the Simple Wikipedia part of the PWKP training set to train an n-gram language model with the SRILM toolkit to learn the probabilities of different n-grams. Then we invoke the GIZA++ aligner using the training simplification pairs. We run GIZA++ with standard settings and we perform no optimization. This results in a phrase table containing phrase pairs from Wikipedia and Simple Wikipedia and their conditional probabilities as assigned by Moses. Finally, we use the Moses decoder to generate simplifications for the sentences in the test set. For each sentence we let the system generate the ten best distinct solutions (or less, if fewer than ten solutions are generated) as ranked by Moses.

Arguably, dissimilarity is a key factor in simplification (and in paraphrasing in general). As output we would like to be able to select fluent sentences that adequately convey the meaning of the original input, yet that contain differences that operationalize the intended simplification. When training our PBMT system on the PWKP data we may assume that the system learns to simplify automatically, yet there is no aspect of the decoder function in Moses that is sensitive to the fact that it should try to be different from the input – Moses may well translate input to unchanged output, as much of our training data consists of partially equal input and output strings.

To expand the functionality of Moses in the intended direction we perform post-hoc re-ranking on the output based on dissimilarity to the input. We do this to select output that is as different as possible from the source sentence, so that it ideally con-

tains multiple simplifications; at the same time, we base our re-ranking on a top-n of output candidates according to Moses, with a small n, to ensure that the quality of the output in terms of fluency and adequacy is also controlled for. Setting $n = 10$, for each source sentence we re-rank the ten best sentences as scored by the decoder according to the Levenshtein Distance (or edit distance) measure (Levenshtein, 1966) at the word level between the input and output sentence, counting the minimum number of edits needed to transform the source string into the target string, where the allowable edit operations are insertion, deletion, and substitution of a single word. In case of a tie in Levenshtein Distance, we select the sequence with the better decoder score. When Moses is unable to generate ten different sentences, we select from the lower number of outputs. Figure 1 displays Levenshtein Distance and Flesch-Kincaid grade level scores for different values of n. We use the Lingua::EN::Fathom module[4] to calculate Flesch-Kincaid grade level scores. The readability score stays more or less the same, indicating no relation between n and readability. The average edit distance starts out at just above 2 when selecting the 1-best output string, and increases roughly until $n = 10$.

2.5 Descriptive statistics

Table 2 displays the average edit distance and the percentage of cases in which no edits were performed for each of the systems and for Simple Wikipedia. We see that the Levenshtein distance between Wikipedia and Simple Wikipedia is the most substantial with an average of 12.3 edits. Given that the average number of tokens is about 25 for Wikipedia and 21 for Simple Wikipedia (cf. Table 1.1), these numbers indicate that the changes in Simple Wikipedia go substantially beyond the average four-word length difference. On average, eight more words are interchanged for other words. About half of the original tokens in the source sentence do not return in the output. Of the three simplification systems, the Zhu system (7.95) and the RevILP (7.18) attain similar edit distances, less substantial than the edits in Simple Wikipedia, but still considerable compared to the baseline word-substitution system (4.26) and PBMT-R (3.08). Our system is clearly conservative in its edits.

System	LD	Perc. no edits
Simple Wikipedia	12.30	3
Word Sub	4.26	0
Zhu	7.95	2
RevILP	7.18	22
PBMT-R	3.08	5

Table 2: Levenshtein Distance and percentage of unaltered output sentences.

On the other hand, we observe some differences in the percentage of cases in which the systems decide to produce a sentence identical to the input. In 22 percent of the cases the RevILP system does not alter the sentence. The other systems make this decision about as often as the gold standard, Simple Wikipedia, where only 3% of sentences remain unchanged. The word-substitution baseline always manages to make at least one change.

3 Evaluation

3.1 Participants

Participants were 46 students of Tilburg University, who participated for partial course credits. All were native speakers of Dutch, and all were proficient in English, having taken a course on Academic English at University level.

3.2 Materials

We use the test set used by Zhu et al. (2010) and Woodsend and Lapata (2011). This test set consists of 100 sentences from articles on English Wikipedia, paired with sentences from corresponding articles in English Simple Wikipedia. We selected only those sentences where every system would perform minimally one edit, because we only want to compare the different systems when they actually generate altered, assumedly simplified output. From this subset we randomly pick 20 source sentences, resulting in 20 clusters of one source sentence and 5 simplified sentences, as generated by humans (Simple Wikipedia) and the four systems.

[4] http://http://search.cpan.org/~kimryan/Lingua-EN-Fathom-1.15/lib/Lingua/EN/Fathom.pm

3.3 Procedure

The participants were told that they participated in the evaluation of a system that could simplify sentences, and that they would see one source sentence and five automatically simplified versions of that sentence. They were not informed of the fact that we evaluated in fact four different systems and the original Simple Wikipedia sentence. Following earlier evaluation studies (Doddington, 2002; Woodsend and Lapata, 2011), we asked participants to evaluate Simplicity, Fluency and Adequacy of the target headlines on a five point Likert scale. Fluency was defined in the instructions as the extent to which a sentence is proper, grammatical English. Adequacy was defined as the extent to which the sentence has the same meaning as the source sentence. Simplicity was defined as the extent to which the sentence was simpler than the original and thus easier to understand. The order in which the clusters had to be judged was randomized and the order of the output of the various systems was randomized as well.

4 Results

4.1 Automatic measures

The results of the automatic measures are displayed in Table 3. In terms of the Flesch-Kincaid grade level score, where lower scores are better, the Zhu system scores best, with 7.86 even lower than Simple Wikipedia (8.57). Increasingly worse Flesch-Kincaid scores are produced by RevILP (8.61) and PBMT-R (13.38), while the word substitution baseline scores worst (14.64). With regard to the BLEU score, where Simple Wikipedia is the reference, the PBMT-R system scores highest with 0.43, followed by the RevILP system (0.42) and the Zhu system (0.38). The word substitution baseline scores lowest with a BLEU score of 0.34.

System	Flesch-Kincaid	BLEU
Simple Wikipedia	8.57	1
Word Sub	14.64	0.34
Zhu	7.86	0.38
RevILP	8.61	0.42
PBMT-R	13.38	0.43

Table 3: Flesch-Kincaid grade level and BLEU scores

4.2 Human judgements

To test for significance we ran repeated measures analyses of variance with system (Simple Wikipedia, PBMT-R, Zhu, RevILP, word-substitution baseline) as the independent variable, and the three individual metrics as well as their combined mean as the dependent variables. Mauchlys test for sphericity was used to test for homogeneity of variance, and when this test was significant we applied a Greenhouse-Geisser correction on the degrees of freedom (for the purpose of readability we report the normal degrees of freedom in these cases). Planned pairwise comparisons were made with the Bonferroni method. Table 4 displays these results.

First, we consider the 3 metrics in isolation, beginning with Fluency. We find that participants rated the Fluency of the simplified sentences from the four systems and Simple Wikipedia differently, $F(4, 180) = 178.436, p < .001, \eta_p^2 = .799$. The word-substitution baseline, Simple Wikipedia and PBMT-R receive the highest scores (3.86, 3.84 and 3.83 respectively) and don't achieve significantly different scores on this dimension. All other pairwise comparisons are significant at $p < .001$. RevILP attains a score of 3.18, while the Zhu system achieves the lowest mean judgement score of 2.59.

Participants also rated the systems significantly differently on the Adequacy scale, $F(4, 180) = 116.509, p < .001, \eta_p^2 = .721$. PBMT-R scores highest (3.71), followed by the word-substitution baseline (3.58), RevILP (3.28), and then by Simple Wikipedia (2.91) and the Zhu system (2.82). Simple Wikipedia and the Zhu system do not differ significantly, and all other pairwise comparisons are significant at $p < .001$. The low score of Simple Wikipedia indicates indirectly that the human editors of Simple Wikipedia texts often choose to deviate quite markedly from the meaning of the original text.

Key to the task of simplification are the human judgements of Simplicity. Participants rated the Simplicity of the output from the four systems and Simple Wikipedia differently, $F(4, 180) = 74.959, p < .001, \eta_p^2 = .625$. Simple Wikipedia scores highest (3.68) and the word substitution baseline scores lowest (2.42). Between them are the RevILP (2.96), Zhu (2.93) and PBMT-R (2.88) sys-

System	Overall	Fluency	Adequacy	Simplicity
Simple Wikipedia	3.46 (0.39)	3.84 (0.46)	2.91 (0.32)	3.68 (0.39)
Word Sub	3.39 (0.43)	3.86 (0.49)	3.58 (0.35)	2.42 (0.48)
Zhu	2.78 (0.45)	2.59 (0.48)	2.82 (0.37)	2.93 (0.50)
RevILP	3.13 (0.36)	3.18 (0.45)	3.28 (0.32)	2.96 (0.39)
PBMT-R	3.47 (0.46)	3.83 (0.49)	3.71 (0.44)	2.88 (0.46)

Table 4: Mean scores assigned by human subjects, with the standard deviation between brackets

	Adequacy	Simplicity	Flesch-Kincaid	BLEU
Fluency	0.45**	0.24*	0.42**	0.26**
Adequacy		-0.19	0.40**	-0.14
Simplicity			-0.45**	0.42**
Flesch-Kincaid				-0.11

Table 5: Pearson correlation between the different dimensions as assigned by humans and the automatic metrics. Scores marked * are significant at $p < .05$ and scores marked ** are significant at $p < .01$

tems, which do not score significantly differently from each other. All other pairwise comparisons are significant at $p < .001$.

Finally we report on a combined score created by averaging over the Fluency, Adequacy and Simplicity scores. Inspection of this score, displayed in the leftmost column of Table 4, reveals that the PBMT-R system and Simple Wikipedia score best (3.47 and 3.46 respectively), followed by the word substitution baseline (3.39), which in turn scores higher than RevILP (3.13) and the Zhu system (2.78). We find that participants rated the systems significantly differently overall, $F(4, 180) = 98.880, p < .001, \eta_p^2 = .687$. All pairwise comparisons were statistically significant ($p < .01$), except the one between the PBMT-R system and Simple Wikipedia.

4.3 Correlations

Table 5 displays the correlations between the scores assigned by humans (Fluency, Adequacy and Simplicity) and the automatic metrics (Flesch-Kincaid and BLEU). We see a significant correlation between Fluency and Adequacy (0.45), as well as between Fluency and Simplicity (0.24). There is a negative significant correlation between Flesch-Kincaid scores and Simplicity (-0.45) while there is a positive significant correlation between Flesch-Kincaid and Adequacy and Fluency. The significant correlations between BLEU and Simplicity (0.42) and Fluency (0.26) are both in the positive direction. There is no significant correlation between BLEU and Adequacy, indicating BLEU's relative weakness in assessing the semantic overlap between input and output. BLEU and Flesch-Kincaid do not show a significant correlation.

5 Discussion

We conclude that a phrase-based machine translation system with added dissimilarity-based re-ranking of the best ten output sentences can successfully be used to perform sentence simplification. Even though the system merely performs phrase-based machine translation and is not specifically geared towards simplification were it not for the dissimilarity-based re-ranking of the output, it performs not significantly differently from state-of-the-art sentence simplification systems in terms of human-judged Simplification. In terms of Fluency and Adequacy our system is judged to perform significantly better. From the relatively low average numbers of edits made by our system we can conclude that our system performs relatively small numbers of changes to the input, that still constitute as sensible simplifications. It does not split sentences (which the Zhu and RevILP systems regularly do); it only rephrases phrases. Yet, it does this better than a word-substitution baseline, which can also be considered a conservative approach; this is reflected in the baseline's high Fluency score (roughly equal to PBMT-R and Simple Wikipedia) and Adequacy score (only slightly worse than PBMT-R).

Wikipedia	the judge ordered that chapman should receive psychiatric treatment in prison and sentenced him to twenty years to life , slightly less than the maximum possible of twenty-five years to life .
Simple Wikipedia	he was sentenced to twenty-five years to life in prison in 1981 .
Word-substitution baseline	the judge ordered that chapman should have psychiatric treatment in prison and sentenced him to twenty years to life , slightly less than the maximum possible of twenty-five years to life .
Zhu	the judge ordered that chapman should get psychiatric treatment . in prison and sentenced him to twenty years to life , less maximum possible of twenty-five years to life .
RevILP	the judge ordered that chapman should will get psychiatric treatment in prison . he sentenced him to twenty years to life to life .
PBMT-R	the judge ordered that chapman should get psychiatric treatment in prison and sentenced him to twenty years to life , a little bit less than the highest possible to twenty-five years to life .

Table 6: Example output

The output of all systems, the original and the simplified version of an example sentence from the PWKP dataset is displayed in Table 6. The Simple Wikipedia sentences illustrate that significant portions of the original sentences may be dropped, and parts of the semantics of the original sentence discarded. We also see the Zhu and RevILP systems resorting to splitting the original sentence in two, leading to better Flesch-Kincaid scores. The word-substitution baseline changes 'receive' in 'have', while the PBMT-R system changes the same 'receive' in 'get', 'slightly' to 'a little bit', and 'maximum' to 'highest'.

In terms of automatic measures we see that the Zhu system scores particularly well on the Flesch-Kincaid metric, while the RevILP system and our PBMT-R system achieve the highest BLEU scores. We believe that for the evaluation of sentence simplification, BLEU is a more appropriate metric than Flesch-Kincaid or a similar readability metric, although it should be noted that BLEU was found only to correlate significantly with Fluency, not with Adequacy. While BLEU and NIST may be used with this in mind, readability metrics should be avoided altogether in our view. Where machine translation evaluation metrics such as BLEU take into account gold references, readability metrics only take into account characteristics of the sentence such as word length and sentence length, and ignore grammaticality or the semantic adequacy of the content of the output sentence, which BLEU is aimed to implicitly approximate by measuring overlap in n-grams. Arguably, readability metrics are best suited to be applied to texts that can be considered grammatical and meaningful, which is not necessarily true for the output of simplification algorithms. A disruptive example that would illustrate this point would be a system that would randomly split original sentences in two or more sequences, achieving considerably lower Flesch-Kincaid scores, yet damaging the grammaticality and semantic coherence of the original text, as is evidenced by the negative correlation for Simplicity and positive correlations for Fluency and Adequacy in Table 5.

In the future we would like to investigate how we can boost the number of edits the system performs, while still producing grammatical and meaning-preserving output. Although the comparison against the Zhu system, which uses syntax-driven machine translation, shows no clear benefit for syntax-based machine translation, it may still be the case that approaches such as Hiero (Chiang et al., 2005) and Joshua (Li et al., 2009), enhanced by dissimilarity-based re-ranking, would improve over our current system. Furthermore, typical simplification operations such as sentence splitting and more radical syntax alterations or even document-level operations such as manipulations of the co-reference structure would be interesting to implement and test

Acknowledgements

We are grateful to Zhemin Zhu and Kristian Woodsend for sharing their data. We would also like to thank the anonymous reviewers for their comments.

References

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604, Morristown, NJ, USA. Association for Computational Linguistics.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 196–205, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yvonne Canning, John Tait, Jackie Archibald, and Ros Crawley. 2000. Cohesive regeneration of syntactically simplified newspaper text. In *Proceedings of ROMAND 2000*, Lausanne.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of English newspaper text to assist aphasic readers. In *AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, Madison, Wisconsin.

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *Proceedings of EACL'99*, Bergen. ACL.

R. Chandrasekar and B. Srinivas. 1997. Automatic rules for text simplification. *Knowledge-Based Systems*, 10:183–190.

Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING'96)*, pages 1041–1044.

David Chiang, Adam Lopez, Nitin Madnani, Christof Monz, Philip Resnik, and Michael Subotin. 2005. The hiero machine translation system: extensions, evaluation, and analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 779–786, Stroudsburg, PA, USA. Association for Computational Linguistics.

Will Coster and David Kauchak. 2011. Learning to simplify sentences using wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 1–9, Portland, Oregon, June. Association for Computational Linguistics.

Walter Daelemans, Jakub Zavrel, Peter Berck, and Steven Gillis. 1996. MBT: A Memory-Based Part of Speech Tagger-Generator. In *Proc. of Fourth Workshop on Very Large Corpora*, pages 14–27. ACL SIGDAT.

Walter Daelemans, Anja Hothker, and Erik Tjong Kim Sang. 2004. Automatic sentence simplification for subtitling in dutch and english. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1045–1048.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, May.

Katja Filippova and Michael Strube. 2008. Sentence fusion via dependency graph compression. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 177–185, Honolulu, Hawaii, October. Association for Computational Linguistics.

Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: A project note. In *Proceedings of the Second International Workshop on Paraphrasing*, pages 9–16, Sapporo, Japan, July. Association for Computational Linguistics.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization – step one: Sentence compression. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI)*, pages 703 – 710, Austin, Texas, USA, July 30 – August 3.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris C. 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 *ACL*. The Association for Computer Linguistics.

V. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren N. G. Thornton, Jonathan Weese, and Omar F. Zaidan. 2009. Joshua: an open source toolkit for parsing-based machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 135–139, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie J. Dorr. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 120–127, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rani Nelken and Stuart M. Shieber. 2006. Towards robust context-sensitive sentence alignment for monolingual corpora. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, Trento, Italy, 3–7 April.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 142–149, Barcelona, Spain, July. Association for Computational Linguistics.

Advaith Siddharthan. 2002. An architecture for a text simplification system. In *Language Engineering Conference*, page 64. IEEE Computer Society.

David A. Smith and Jason Eisner. 2006. Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. In *Proceedings of the HLT-NAACL Workshop on Statistical Machine Translation*, pages 23–30, New York, June.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *In Proc. Int. Conf. on Spoken Language Processing*, pages 901–904, Denver, Colorado.

D. Vickrey and D. Koller. 2008. Sentence simplification for semantic role labeling. In *Proceedings of the 46th Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Willian Massami Watanabe, Arnaldo Candido Junior, Vincius Rodriguez de Uzłda, Renata Pontin de Mattos Fortes, Thiago Alexandre Salgueiro Pardo, and Sandra M. Alusio. 2009. Facilita: reading assistance for low-literacy readers. In Brad Mehlenbacher, Aristidis Protopsaltis, Ashley Williams, and Shaun Slattery, editors, *SIGDOC*, pages 29–36. ACM.

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

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2010. Paraphrase generation as monolingual translation: data and evaluation. In *Proceedings of the 6th International Natural Language Generation Conference*, INLG '10, pages 203–207, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 523–530, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Proceedings of the NAACL*, pages 365–368.

Shiqi Zhao, Xiang Lan, Ting Liu, and Sheng Li. 2009. Application-driven statistical paraphrase generation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 834–842, Stroudsburg, PA, USA. Association 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 2010)*, pages 1353–1361, Beijing, China, August. Coling 2010 Organizing Committee.