Optimizing Statistical Machine Translation for Text Simplification

Wei Xu[1], Courtney Napoles[2], Ellie Pavlick[1], Quanze Chen[1] and **Chris Callison-Burch[1]**
[1] Computer and Information Science Department
University of Pennsylvania
{xwe, epavlick, cquanze, ccb}@seas.upenn.edu
[2] Department of Computer Science
Johns Hopkins University
courtneyn@jhu.edu

Abstract

Most recent sentence simplification systems use basic machine translation models to learn lexical and syntactic paraphrases from a manually simplified parallel corpus. These methods are limited by the quality and quantity of manually simplified corpora, which are expensive to build. In this paper, we conduct an in-depth adaptation of statistical machine translation to perform text simplification, taking advantage of large-scale paraphrases learned from bilingual texts and a small amount of manual simplifications with multiple references. Our work is the first to design automatic metrics that are effective for tuning and evaluating simplification systems, which will facilitate iterative development for this task.[1]

1 Introduction

The goal of text simplification is to rewrite an input text so that the output is more readable. Text simplification has applications for reducing input complexity for natural language processing (Siddharthan et al., 2004; Miwa et al., 2010; Chen et al., 2012b) and providing reading aids for people with limited language skills (Petersen and Ostendorf, 2007; Watanabe et al., 2009; Allen, 2009; De Belder and Moens, 2010; Siddharthan and Katsos, 2010) or language impairments such as dyslexia (Rello et al., 2013), autism (Evans et al., 2014), and aphasia (Carroll et al., 1999).

It is widely accepted that sentence simplification can be implemented by three major types of operations: *splitting*, *deletion* and *paraphrasing* (Feng, 2008; Narayan and Gardent, 2014). The splitting operation decomposes a long sentence into a sequence of shorter sentences. Deletion removes less important parts of a sentence. The paraphrasing operation includes reordering, lexical substitutions and syntactic transformations. While sentence splitting (Siddharthan, 2006; Petersen and Ostendorf, 2007; Narayan and Gardent, 2014; Angrosh et al., 2014) and deletion (Knight and Marcu 2002; Clarke and Lapata 2006; Filippova and Strube 2008; and others) have been intensively studied, there has been considerably *less* research on developing new paraphrasing models for text simplification — most previous work has used off-the-shelf statistical machine translation (SMT) technology and achieved reasonable results (Coster and Kauchak, 2011a,b; Wubben et al., 2012; Štajner et al., 2015). However, they have either treated the SMT technology as a black box (Coster and Kauchak, 2011a,b; Narayan and Gardent, 2014; Angrosh et al., 2014; Štajner et al., 2015) or they have been limited to modifying only one aspect of it, such as the translation model (Zhu et al., 2010; Woodsend and Lapata, 2011) or the reranking component (Wubben et al., 2012).

We treat simplification as a monolingual text-to-text generation problem. We use a large-scale paraphrase database in combination with the machinery from statistical machine translation, following Ganitkevitch et al. (2011) who employed the same methodology to shorten sentences. Our methodology poses text simplification as a paraphrasing problem. Given an input text, rewrite it subject to the constraints that the output should be simpler than the

[1] We will make our data and system publicly available.

input, while preserving as much meaning of the input as possible, and remaining well-formed English. Our approach is primarily focused on lexical simplification (rewriting words or phrases with simpler versions), and to a lesser extent on syntactic rewrite rules that simplify the input. It largely ignores the important subtasks of sentence splitting and deletion. Our focus on lexical simplification does not affect the generality of the presented work, since deletion or sentence splitting could be applied as pre- or post-processing steps.

In this paper, we present a complete adaptation of a syntax-based machine translation framework to perform simplification. Going beyond previous work, we make direct modifications to four key components in the SMT pipeline: 1) two novel simplification-specific tunable metrics; 2) large-scale paraphrase rules from bilingual pivoting; 3) rich rule-level simplification features; and 4) multiple reference simplifications collected via crowdsourcing. In particular, we report the first study that shows promising correlations of automatic metrics with human evaluation. Our work answers the call made in a recent TACL paper (Xu et al., 2015) to address problems in current simplification research — we amend human evaluation criteria, develop automatic metrics, and generate an improved multiple reference dataset.

2 Background

Xu et al. (2015) laid out a series of problems that are present in current text simplification research, and argued that we should deviate from the previous state-of-the-art benchmarking setup.

First, the Simple English pedia data has dominated simplification research since 2010 (Zhu et al., 2010; Siddharthan, 2014), and is used together with Standard English Wikipedia to create parallel text to train MT-based simplification systems. However, the parallel Wikipedia simplification corpus contains a large proportion of inadequate (not much simpler) or inaccurate (not aligned or only partially aligned) simplifications (Xu et al., 2015; Hwang et al., 2015). It is one of the leading reasons that existing simplification systems struggle to generate simplifying paraphrases and leave the input sentences unchanged (Wubben et al., 2012). Previously researchers attempted some quick fixes by adding phrasal deletion rules (Coster and Kauchak, 2011a) or reranking n-best outputs based on their dissimilarity to the input (Wubben et al., 2012). In contrast, we exploit data with improved quality and enlarged quantity, namely, large-scale paraphrase rules automatically derived from bilingual corpora and a small amount of manual simplification data with multiple references for tuning parameters. We then systematically design new tuning metrics and rich simplification-specific features into a syntactic machine translation model to enforce optimization towards simplicity. This approach achieves better simplification performance without relying on a manually simplified corpus to learn paraphrase rules, which is important given the fact that Simple Wikipedia is only available for English.

Second, previous evaluation used in the simplification literature is uninformative and incomparable across models due to the complications between three different operations of paraphrasing, deletion, and splitting. This, combined with the unreliable quality of Simple Wikipedia as a gold reference for evaluation, has been the bottleneck for developing automatic metrics. There exist only a few studies (Wubben et al., 2012; Štajner et al., 2014) on automatic simplification evaluation using existing MT metrics which show limited correlation with human assessments. In this paper, we restrict ourselves to lexical simplification, where we believe MT-derived evaluation metrics can best be deployed. Our newly proposed metric is the first automatic metric that shows reasonable correlation with human evaluation on the text simplification task. We also introduce multiple references to make automatic evaluation feasible.

The most related work to ours is that of Ganitkevitch et al. (2013) on sentence compression, in which compression of word and sentence lengths can be more straightforwardly implemented in features and the objective function in the SMT framework. Our work is also related to other tunable metrics designed to be very simple and light-weight to ensure fast repeated computation for tuning bilingual translation models (Liu et al., 2010; Chen et al., 2012a). To the best of our knowledge, no tunable metric has been attempted for simplification, except for BLEU. Nor do any evaluation metrics exist for

simplification, although there are several designed for other text-to-text generation tasks: grammatical error correction (Napoles et al., 2015; Felice and Briscoe, 2015; Dahlmeier and Ng, 2012), paraphrase generation (Chen and Dolan, 2011; Xu et al., 2012; Sun and Zhou, 2012), and conversation generation (Galley et al., 2015). Another line of related work is lexical simplification that focuses on finding simpler synonyms of a given complex word (Yatskar et al., 2010; Biran et al., 2011; Specia et al., 2012; Horn et al., 2014).

3 Adapting Machine Translation for Simplification

We adapt the machinery of statistical machine translation (SMT) to the task of text simplification by making changes in the following four key components:

3.1 Simplification-specific Objective Functions

In the statistical machine translation framework, one crucial element is to design automatic evaluation metrics to be used as training objectives. Training algorithms, such as MERT (Och, 2003) or PRO (Hopkins and May, 2011), then directly optimize the model parameters such that the end-to-end simplification quality is optimal. Unfortunately, previous work on text simplification has only used BLEU for tuning, which is insufficient as we show empirically in Section 4. We propose two new light-weight metrics instead: one that explicitly measures readability and the other implicitly measures it by comparing against the input and references.

Unlike machine translation metrics which do not compare against the (foreign) input sentence, it is necessary to compare simplification system outputs against the inputs to assess readability changes. It is also important to keep tunable metrics as simple as possible, since they are repeatedly computed during the tuning process for hundreds of thousands candidate outputs.

FKBLEU

Our first metric combines a previously proposed metric for paraphrase generation, $iBLEU$ (Sun and Zhou, 2012), and the widely used readability metric, Flesch-Kincaid Index (Kincaid et al., 1975). iBLEU is an extension of the BLEU metric to measure diversity as well as adequacy of the generated paraphrase output. Given a candidate sentence O, human references R and input text I, iBLEU is defined as:

$$\begin{aligned} \text{iBLEU} &= \alpha \times \text{BLEU}(O, R) \\ &\quad -(1-\alpha) \times \text{BLEU}(O, I). \end{aligned} \tag{1}$$

where α is a parameter taking balance between adequacy and dissimilarity, and set to 0.9 empirically as suggested by Sun and Zhou (2012).

Since the text simplification task aims at improving readability. Thus, we include the Flesch-Kincaid Index (*FK*) which estimates the readability of text using cognitively motivated features (Kincaid et al., 1975):

$$\begin{aligned} \text{FK} &= 0.39 \times \left(\frac{\#\text{words}}{\#\text{sentences}} \right) \\ &\quad +11.8 \times \left(\frac{\#\text{syllables}}{\#\text{words}} \right) - 15.59 \end{aligned} \tag{2}$$

with a lower value indicating higher readability.[2] We adapt FK to score individual sentences and change it so that it counts punctuation tokens as well as word, and assigning each punctuation token as one syllable. This prevents it from arbitrarily deleting punctuation.

FK measures readability *assuming that the text is well-formed*, and therefore is insufficient alone as a metric for generating or evaluating automatically generated sentences. Combining FK and iBLEU captures both a measure of readability and adequacy. The resulting objective function, ***FKBLEU***, is defined as a geometric mean of the iBLEU and the FK difference between input and output sentences:

$$\begin{aligned} \text{FKBLEU} &= \text{iBLEU}(I, R, O)^{1/2} \\ &\quad \times \text{FKdiff}(I, O)^{1/2} \\ \text{FKdiff} &= \text{sigmoid}(\text{FK}(O) - \text{FK}(I)). \end{aligned} \tag{3}$$

Sentences with higher FKBLEU values are better simplification with higher readability.

[2] The FK coefficients were derived via multiple regression applied to the reading compression test scores of 531 Navy personnel reading training manuals. These values are typically used unmodified, as we do here.

SARI

We design a second new metric $SARI$ that principally compares **s**ystem output **a**gainst **r**eferences and against the **i**nput sentence. It explicitly measures the goodness of words that are added, deleted and kept by the systems (Figure 1).

We reward addition operations, where system output O that was not in the input I but that occurred in any of the references R, i.e. $O \cap \overline{I} \cap R$. We define n-gram precision $p(n)$ and recall $r(n)$ for addition operations as follows:[3]

$$p_{add}(n) = \frac{\sum_{g \in O} \min\left(\#_g(O \cap \overline{I}), \#_g(R)\right)}{\sum_{g \in O} \#_g(O \cap \overline{I})}$$

$$r_{add}(n) = \frac{\sum_{g \in O} \min\left(\#_g(O \cap \overline{I}), \#_g(R)\right)}{\sum_{g \in O} \#_g(O \cap \overline{R})}$$

where $\#_g(\cdot)$ is a binary indicator of occurrence of n-grams g in a given set (and is a fractional indicator in some later formulas) and

$$\#_g(O \cap \overline{I}) = \max(\#_g(O) - \#_g(I), 0)$$
$$\#_g(O \cap \overline{R}) = \max(\#_g(O) - \#_g(R), 0)$$

Therefore, in the example below, the addition of unigram *now* is rewarded in both $p_{add}(n)$ and $r_{add}(n)$, while the addition of *you* in OUTPUT-1 is penalized in $p_{add}(n)$:

INPUT: *About 95 species are currently accepted .*
REF-1: *About 95 species are currently known .*
REF-2: *About 95 species are now accepted .*
REF-3: *95 species are now accepted .*
OUTPUT-1: *About 95 you now get in .*
OUTPUT-2: *About 95 species are now accepted .*
OUTPUT-3: *About 95 species are currently agreed.*

The corresponding SARI scores of the three toy outputs are 0.2683, 0.7594, 0.5890, which match with intuitions about their quality. To put in perspective, the BLEU scores are 0.1562, 0.6435, 0.6435 respectively. BLEU fails to distinguish between OUTPUT-2 and the OUTPUT-3 because matching any one of references is credited the same. Not all the references are necessarily complete simplifications, e.g. REF-1 doesn't simplify the word *currently*, which gives BLEU too much latitude for matching the input.

[3] In the rare case when the denominator is zero in calculating precision p or recall r, we simply set the value of p and r to 0.

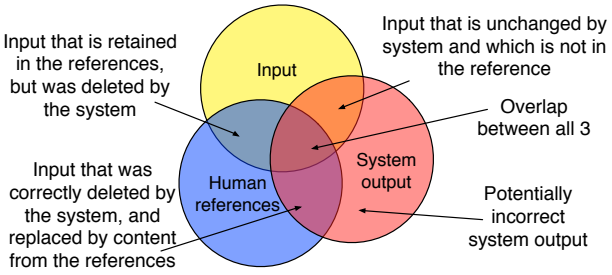

Figure 1: Metrics that evaluate the output of monolingual text-to-text generation systems can compare system output against references and against the input sentence, unlike in MT metrics which do not compare against the (foreign) input sentence. The different regions of this Venn diagram are treated differently with our SARI metric.

Words that are retained in both the system output and references should be rewarded. When multiple references are used, the number of references in which an n-gram was retained matters. It takes into account that some words/phrases are considered simple and are unnecessary (but still encouraged) to be simplified. We use R' to mark the n-gram counts over R with fractions, e.g. if a unigram (*about* in above example) occurs in 2 out of the total r references, then its count is weighted by $2/r$ in computation of precision and recall:

$$p_{keep}(n) = \frac{\sum_{g \in I} \min\left(\#_g(I \cap O), \#_g(I \cap R')\right)}{\sum_{g \in I} \#_g(I \cap O)}$$

$$r_{keep}(n) = \frac{\sum_{g \in I} \min\left(\#_g(I \cap O), \#_g(I \cap R')\right)}{\sum_{g \in I} \#_g(I \cap R')}$$

where

$$\#_g(I \cap O) = \min\left(\#_g(I), \#_g(O)\right)$$
$$\#_g(I \cap R') = \min\left(\#_g(I), \#_g(R)/r\right)$$

For deletion, we only use precision because over-deleting hurts readability much more significantly than not deleting:

$$p_{del}(n) = \frac{\sum_{g \in I} \min\left(\#_g(I \cap \overline{O}), \#_g(I \cap \overline{R'})\right)}{\sum_{g \in I} \#_g(I \cap \overline{O})}$$

where

$$\#_g(I \cap \overline{O}) = \max\left(\#_g(I) - \#_g(O), 0\right)$$
$$\#_g(I \cap \overline{R'}) = \max\left(\#_g(I) - \#_g(R)/r, 0\right)$$

Lexical	[RB]	solely	→	only
	[NN]	objective	→	goal
	[JJ]	undue	→	unnecessary
Phrasal	[VP]	accomplished	→	carried out
	[VP/PP]	make a significant contribution	→	contribute greatly
	[VP/S]	is generally acknowledged that	→	is widely accepted that
Syntactic	[NP/VP]	the manner in which NN	→	the way NN
	[NP]	NNP 's population	→	the people of NNP
	[NP]	NNP 's JJ legislation	→	the JJ law of NNP

Table 1: Example paraphrase rules in the Paraphrase Database (PPDB) that result in simplifications of the input. The rules are SCFG rules where uppercase indicates non-terminal symbols. Non-terminal can be complex symbols like VP/S which indicates that the rule forms a verb phrase missing a sentence to its right. The final syntactic rule both simplifies and reorders the input phrase.

The precision of what is kept also reflects the sufficiency of deletions. The n-gram counts are also weighted in $\overline{R'}$ to compensate n-grams, such as *currently* in the example, that are not considered as *must be simplified* by human editors.

Together, in SARI, we use arithmetic average of n-gram precisions $P_{operation}$ and recalls $R_{operation}$:

$$SARI = d_1 F_{add} + d_2 F_{keep} + d_3 P_{del} \quad (4)$$

where $d_1 = d_2 = d_3 = 1/3$ and

$$P_{operation} = \frac{1}{k} \sum_{n=[1,\ldots,k]} p_{operation}(n)$$

$$R_{operation} = \frac{1}{k} \sum_{n=[1,\ldots,k]} r_{operation}(n)$$

$$F_{operation} = \frac{2 \times P_{operation} \times R_{operation}}{P_{operation} + R_{operation}}$$

$$operation \in [del, keep, add]$$

where k is the highest n-gram order and set to 4 in our experiments.

3.2 Incorporating Large-Scale Paraphrase Rules

Another challenge for text simplification is generating an ample set of rewrite rules that potentially simplify an input sentence. Most early work has relied on either hand-crafted rules (Chandrasekar et al., 1996; Carroll et al., 1999; Siddharthan, 2006; Vickrey and Koller, 2008) or dictionaries like WordNet (Devlin et al., 1999; Kaji et al., 2002; Inui et al., 2003). Other more recent studies have relied on the parallel Normal-Simple Wikipedia Corpus to automatically extract rewrite rules. This technique does manage to learn a small number of transformations that simplify. However, we argue that because the size of the Normal-Simple Wikipedia parallel corpus is quite small (108k sentence pairs with 2 million words), the diversity and coverage of patterns that can be learned is actually quite limited.

In this paper we will leverage the large-scale Paraphrase Database (PPDB) (Ganitkevitch et al., 2013; Pavlick et al., 2015)[4] as a rich source of lexical, phrasal and syntactic simplification operations. The PPDB is represented as a synchronous context-free grammar (SCFG), which is commonly used as the formalism for syntax-based machine translation. It is created by extracting English paraphrases from bilingual parallel corpora using a technique called "bilingual pivoting" (Bannard and Callison-Burch, 2005). Table 1 shows some example paraphrase rules in the PPDB.

PPDB employs 1000 times more data (106 million sentence pairs with 2 billion words) than the Normal-Simple Wikipedia parallel corpus. The key differences between the paraphrase rules from PPDB and the transformations learned by the naive application of SMT to the Normal-Simple Wikipedia parallel corpus, are that the PPDB paraphrases are much more diverse. For example, PPDB contains 214 paraphrases for *ancient* including *antique, ancestral, old, age-old, archeological, for-*

[4] http://paraphrase.org

mer, antiquated, longstanding, archaic, centuries-old, and so on. However, there is nothing inherent in the rule extraction process to say which of the PPDB paraphrases are simplifications.

One way of using PPDB for simplification would be to simply discard any of its rules which did not result in a simplified output. Instead of filtering the rules to contain only simplifying operations, we model the task by incorporating rich features into each rule and let SMT advances in decoding and optimization to determine how well a rule simplifies an input phrase.

3.3 Simplification-specific Features for Paraphrase Rules

Designing good features is an essential aspect of modeling. Each paraphrase rule has a vector $\vec{\varphi} = \{\varphi_1 ... \varphi_N\}$ of feature functions that are combined in a linear model to obtain a single score w for each rule application:

$$w = -\sum_{i=1}^{N} \lambda_i \log \varphi_i. \quad (5)$$

In SMT, typical feature functions are phrase translation probabilities, word-for-word lexical translation probabilities, a rule application penalty (which governs whether the system prefers fewer longer phrases or a greater number of shorter phrases), and a language model probability. Together these features are what the model uses to distinguish between good and bad translations. For monolingual translation tasks, previous research suggests that features like paraphrase probability and distributional similarity are potentially helpful in picking out good paraphrases (Chan et al., 2011) and for text-to-text generation (Ganitkevitch et al., 2012b). While these two features quantify how good a paraphrase rule is *in general*, they do not indicate how good the rule is for a specific task, i.e. simplification (Pavlick and Nenkova, 2015).

We use all the features that were distributed with PPDB 1.0 and add new features for simplification purposes:[5] length in characters, length in words, number of syllables, language model scores, and fraction of common English words in each rule. These features are computed both sides of a paraphrase pattern, the word with the maximum number of syllables on each side and the difference between the two sides, when it is applicable. We use language models built from the Gigaword corpus and the Simple Wikipedia corpus collected by Kauchak (2013). We also use a list of top 3000 US English words compiled by Paul and Bernice Noll.[6]

3.4 Creating Multiple References

Like with machine translation, where there are many equally good translations, in simplification there may be several ways of simplifying a sentence. Most previous work on text simplification only uses a single reference simplification, often from parallel Wikipedia simplification corpus. This is undesirable since recent studies (Xu et al., 2015; Amancio and Specia, 2014; Hwang et al., 2015; Štajner et al., 2015) suggest that the Simple Wikipedia contains a large proportion of inadequate (not much simpler) or inaccurate (not aligned or only partially aligned) simplifications.

In this study, we collect multiple human reference simplifications that focus on simplification by paraphrasing rather than deletion or splitting. We first selected the Simple-Normal sentence pairs of similar length (<= 20% differences in number of tokens) from the Parallel Wikipedia Simplification (PWKP) corpus (Zhu et al., 2010) that are more likely to be paraphrase-only simplifications. We then asked 8 workers on Amazon Mechanical Turk to rewrite a selected sentence from Normal Wikipedia (a subset of PWKP) into a simpler version while preserving its meaning, without losing any information or splitting sentence. We removed bad workers by manual inspection on worker's first several submissions on the basis of a recent study (Gao et al., 2015) on crowdsourcing translation that suggests Turkers' performance stay consistent over time and can be reliably predicted by their first few translations.

In total, we collected 8 reference simplifications for 2350 sentences, and randomly split them into 2000 sentences for tuning, 350 for evaluation. Many crowdsourcing workers were able to provide simplifications of good quality and diversity (see Table 5

[5] We will release the data with detailed description for each feature.

[6] http://www.manythings.org/vocabulary/lists/l/noll-about.php

for manual quality evaluation and Table 2 for an example). Having multiple references allows us to develop automatic metrics similar to BLEU to take advantage of the variation across many people's simplifications. We leave more in-depth investigations on crowdsourcing simplification (Pellow and Eskenazi, 2014a,b) for future work.

3.5 Tuning Parameters

Like in statistical machine translation, we set the weights of the linear model λ_i so that the system's output is optimized with respect to the automatic evaluation metric on the 2000 sentence development set. This process is known as minimum error rate training (Och, 2003). We use the PRO algorithm in the open-source Joshua toolkit (Ganitkevitch et al., 2012a; Post et al., 2013) to perform pairwise ranking optimization (Hopkins and May, 2011).

4 Experiments and Analyses

We implemented all the proposed adaptations into the open source syntactic machine translation decoder Joshua (Post et al., 2013),[7] and conducted the experiments with PPDB and the dataset of 2350 sentences collected in Section 3.4. Most recent end-to-end sentence simplification systems either use a basic phrase-based MT model trained on parallel Wikipedia data using the Moses decoder (Štajner et al., 2015) or conjunct paraphrasing operation together with deletion and splitting. One of the best systems is PBMT-R by Wubben et al. (2012), which reranks Moses' n-best outputs based on their dissimilarity to the input to promote simplification. We build a second baseline by using BLEU as the tuning metric in our adapted MT framework for comparison. We conduct both human and automatic evaluation to demonstrate the advantage of the proposed simplification systems. We also show the effectiveness of the two new metrics in tuning and automatic evaluation.

4.1 Qualitative Analysis

Table 2 shows an representative example of the simplification results. The PBMT-R model failed to learn any good substitutions to the word "able-bodied" from the manually simplified corpora of limited size. In contrast, our proposed method can make use of more paraphrases learned from the more abundant bilingual texts. It improves method applicability to languages other than English, for which no simpler version of pedia is available.

[7] http://joshua-decoder.org/

Paraphrase Rule	Trans. Model Score
principal → key	4.515
principal → main	**4.514**
principal → major	4.358
principal → chief	3.205
principal → core	3.025
principal → principal	2.885
principal → top	2.600
principal → senior	2.480
principal → lead	2.377
principal → primary	2.171
principal → prime	1.432
principal → keynote	-0.795
able-bodied → valid	6.435
able-bodied → sound	**5.838**
able-bodied → healthy	4.446
able-bodied → able-bodied	3.372
able-bodied → job-ready	1.611
able-bodied → employable	-0.363
able-bodied → non-disabled	-2.207

Table 3: Qualitative analysis of candidate paraphrases ranked by the translation model in SBMT (PPDB + SARI), showing that the model is optimized towards simplicity in addition to the correctness of paraphrases. The final simplifications (in bold) are chosen in conjunction with the language model to fit the context and further bias towards more common n-grams.

Our proposed approach also provides an intuitive way to inspect the ranking of candidate paraphrases in the translation model. This is done by scoring each rule in PPDB by Equation 5 using the weights optimized in the tuning process, as in Table 3. It shows that our proposed method is capable of capturing the notion of simplicity using a small amount of parallel tuning data. It correctly ranks *key* and *main* as good simplifications for *principal*. It's choices are not always perfect as it prefers *sound* over *healthy* for *able-bodied*. The final simplification outputs are generated according to both the translation model and the language model to take into account of context and further bias towards more common n-grams.

	Sentence
Normal Wikipedia	Jeddah is the principal gateway to Mecca, Islam's holiest city, which able-bodied Muslims are required to visit at least once in their lifetime.
Simple Wikipedia	Jeddah is the main gateway to Mecca, the holiest city of Islam, where able-bodied Muslims must go to at least once in a lifetime.
Mechanical Turk #1	Jeddah is the main entrance to Mecca, the holiest city in Islam, which all healthy Muslims need to visit at least once in their life.
Mechanical Turk #2	Jeddah is the main entrance to Mecca, Islam's holiest city, which pure Muslims are required to visit at least once in their lifetime.
PBMT-R (Wubben et al., 2012)	Jeddah is the main gateway to Mecca, Islam 's holiest city, which Muslims are required of Muslims at least once in their lifetime.
SBMT (PPDB + BLEU)	Jeddah is the main door to Mecca, Islam holiest city, which sound Muslims are to go to at least in life.
SBMT (PPDB + FKBLEU)	Jeddah is the main gateway to Mecca, Islam's holiest city, which sound Muslims must visit at least once in life.
SBMT (PPDB + SARI)	Jeddah is the main gateway to Mecca, Islam's holiest city, which sound Muslims have to visit at least once in their life.

Table 2: Example human reference simplifications and automatic simplification system outputs. Previous work (e.g. PBMT-R) that learns paraphrase rules from the parallel Normal-Simple Wikipedia corpus is prone to sentence alignment and word alignment errors, and thus erroneous rules such as "able-bodied → Muslims" in this case.

4.2 Quantitative Evaluation of Simplification Systems

For the human evaluation, participants were shown the original English Wikipedia sentence as a reference, and asked to judge a set of simplifications that were displayed in random order. They evaluated a simplification from each system, the Simple pedia version of the reference, and a Turker simplification. Judges rated each simplification on two 5-point scales of meaning retention and grammaticality (0 is the worst and 4 is the best). We also ask participants to rate Simplicity Gain (Simplicity+), by counting how many successful lexical or syntactic paraphrases occurred in the simplification. We found this makes the judgement clearer, easier and more informative than rating the simplicity directly on 5-point scale, since the original sentences have very different readability levels to start with. More importantly, using simplicity gain avoids over-punishment on any errors, which are already penalized for poor meaning retention and grammaticality, and thus reduce the bias towards very conservative models. We collect judgements on these three criteria from 5 different annotators and report the average scores.

Table 5 shows that our best system, a syntactic-based MT system (SBMT) using PPDB as the source of paraphrase rules and tuning towards the SARI metric, achieves better performance in all three simplification measurements than the state-of-the-art system PBMT-R. The relatively small numbers of simplicity gain, even with only two human references (Simple Wikipedia and Mechanical Turk), clearly show the major challenge of simplification, which is the need of not only generating paraphrases but also ensuring the generated paraphrases are simpler while fitting the contexts. Although many researchers have noticed this difficulty, PBMT-R is one of the few that tried to address it by promoting outputs that are dissimilar to the input. Our best system is able to make more effective paraphrases (better Simplicity+) while introducing less errors (better Grammar and Meaning).

Table 6 shows the computation time for different metrics. SARI is only slightly slower than BLEU but achieves much better simplification quality.

	Time (milliseconds)
BLEU	0.12540908
FKBLEU	1.2527733
SARI	0.15506646

Table 6: Average computation time of different metrics per candidate sentence.

	FK	BLEU	iBLEU	FKBLEU	SARI
Normal Wikipedia	12.88	99.05	78.41	62.48	26.05
Simple pedia	11.25	66.75	53.53	61.75	38.42
Mechanical Turk	10.80	100.0	74.31	73.60	43.71
PBMT-R (Wubben et al., 2012)	11.10	63.12	48.91	59.00	33.77
SBMT (PPDB + BLEU-8ref)	12.88	99.05	78.41	62.48	26.05
SBMT (PPDB + FKBLEU-8ref)	10.75	74.48	58.10	66.68	34.18
SBMT (PPDB + SARI-8ref)	10.90	72.36	58.15	66.57	37.91

Table 4: Automatic evaluation of different simplification systems. Most systems achieves similar FK readability scores as human. The SARI metric ranks all 5 different systems and 3 human references in the same order as human assessment. Tuning towards BLEU with all 8 references results in identical transformation (same as **Normal Wikipedia**), as this can get a near-perfect BLEU score of 99.05 (out of 100).

	Grammar	Meaning	Simplicity+	#tokens	#chars	Edit Dist.
Normal Wikipedia	4.00	4.00	0.00	23	125	0.00
Simple Wikipedia	3.72	3.24	1.03	22	116	6.69
Mechanical Turk	3.70	3.36	1.35	19	104	8.25
PBMT-R (Wubben et al., 2012)	3.18	2.83	0.47	20	108	5.96
SBMT (PPDB + BLEU-8ref)	4.00	4.00	0.00	23	125	0.00
SBMT (PPDB + FKBLEU-8ref)	3.30	3.05	0.48	21	107	4.03
SBMT (PPDB + SARI-8ref)	**3.50**	**3.16**	**0.65**	23	118	3.98

Table 5: Human evaluation (Grammar, Meaning, Simplicity+) and basic statistics of our proposed systems (SBMTs) and baselines. PBMT-R is an reimplementation of the state-of-the-art system by Wubben et al. (2012). Newly proposed metrics FKBLEU and SARI show advantages for tuning.

Spearman's ρ	ref.	Grammar	Meaning	Simplicity+
FK	none	- 0.002 ($\approx$ .976)	0.136 ($<$.010)	0.147 ($<$.010)
BLEU	single	0.366 ($<$.001)	0.459 ($<$.001)	**0.151** ($<$.005)
BLEU	multiple	**0.589** ($<$.001)	**0.701** ($<$.001)	0.111 ($<$.050)
iBLEU	single	0.313 ($<$.001)	0.397 ($<$.001)	0.149 ($<$.005)
iBLEU	multiple	0.492 ($<$.001)	0.609 ($<$.001)	0.141 ($<$.010)
FKBLEU	multiple	0.349 ($<$.001)	0.410 ($<$.001)	0.235 ($<$.001)
SARI	multiple	**0.342** ($<$.001)	**0.397** ($<$.001)	**0.343** ($<$.001)

Table 7: Correlations (and p-values) of metrics against the human ratings at sentence-level. In this work, we propose to use multiple (eight) references and two new metrics: FKBLEU and SARI. For all three criteria of simplification quality, SARI correlates reasonably with human judgments. In contrast, previous work use only single reference. Existing metrics BLEU and iBLEU show higher correlations on grammaticality and meaning preservation using multiple references, but fail to measure the most important aspect of simplification – simplicity.

4.3 Correlation of Automatic Metrics with Human Judgements

Table 7 shows the correlation of automatic metrics with human judgement. There are several interesting observations. First, the correlation between automatic metrics with human judgement of grammaticality and meaning preservation is higher than any reported before (Wubben et al., 2012; Štajner et al., 2014). It validates our argument that constraining simplification to only paraphrasing reduces the complication from deletion and splitting, and thus makes automatic evaluation more feasible. Using multiple references further improves the correlations. Second, same as noted in previous work (Wubben et al., 2012; Štajner et al., 2014), none of the existing metrics demonstrate any significant correlation with the simplicity scores rated by humans. However, simplicity is essential in measuring the goodness of sim-

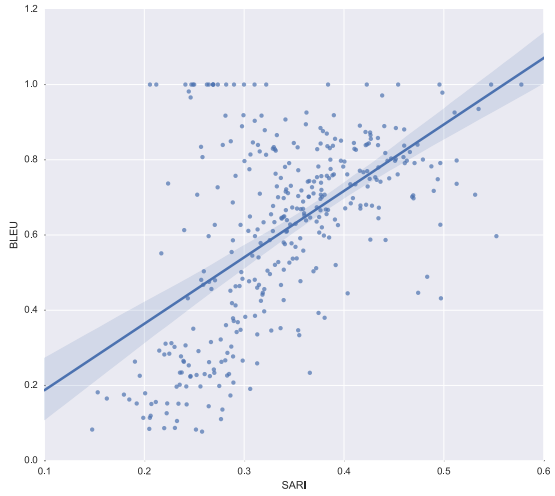

Figure 2: A scatter plot of BLEU scores vs. SARI scores for the individual sentences in our test set. The metrics' scores for many sentences substantially diverge. Few of the sentences that scored perfectly in BLEU receive a high score from SARI.

plification. Third, our two new metrics, FKBLEU and SARI, achieve a much better correlation with humans in simplicity judgement while still capturing the notion of grammaticality and meaning preservation. This explains why they are more suitable than BLEU to be used in training the simplification models. In particular, SARI provides a balanced and integrative measurement of system performance that can assist iterative development. Till today, developing advanced simplification systems has been a difficult and time-consuming process, since it is impractical to run new human evaluation every time a new model is built or parameters are adjusted.

4.4 Why Does BLEU Correlate Strongly with Meaning/Grammar, and SARI with Simplicity?

Here we look more deeply at the correlations of BLEU and SARI with human judgments. Our SARI metric has highest correlation with human judgments of simplicity, but BLEU exhibits higher correlations on grammaticality and meaning preservation.

BLEU was designed to evaluate translations systems. It measures the n-gram precision of a system's output against one or more references. BLEU ignores recall (and compensates for this with its brevity penalty). BLEU prefers an output that is not too short and contains only n-grams that appear in any reference. The role of multiple references in BLEU is to capture allowable variations in translation quality.

When applied to monolingual tasks like simplification, BLEU does not take into account anything about the differences between the input and the references. In contrast, SARI takes into account both precision and recall, by looking at the difference between the references and the input sentence. In this work, we use multiple references to capture many different ways of simplifying the input.

Unlike bilingual translation, the more references created for the monolingual simplification task the more n-grams of the original input will be included in the references. That means, with more references, outputs that are close or identical to the input will get high BLEU. Outputs with few changes also receive high Grammar/Meaning scores from human judges; but these do not necessarily get high SARI score **nor** are they good simplifications.

BLEU therefore tends to favor conservative systems that do not make many changes, while SARI penalizes them. This can be seen in Figure 2 where sentences with a BLEU score of 1.0, receive a range of scores from SARI.

The scatter plots in Figure 3 further illustrate the above analysis. These plots emphasize the correlation of high human scores for meaning/grammar for conservative systems that make few changes (which BLEU rewards, but SARI does not). The tradeoff is that conservative outputs with few or no changes do not result in increased simplicity. SARI correctly rewards systems that make changes that simplify the input.

5 Conclusions and Future Work

In this paper, we presented an effective adaptation of statistical machine translation techniques. We find the approach promising in suggesting two new directions: designing tunable metrics that correlate with humans and using simplicity-enriched paraphrase rules derived from larger data than the Normal-Simple Wikipedia dataset. We will make our system and new dataset with multiple references publicly available. For future work, we think it might be possible to design a universal metric that

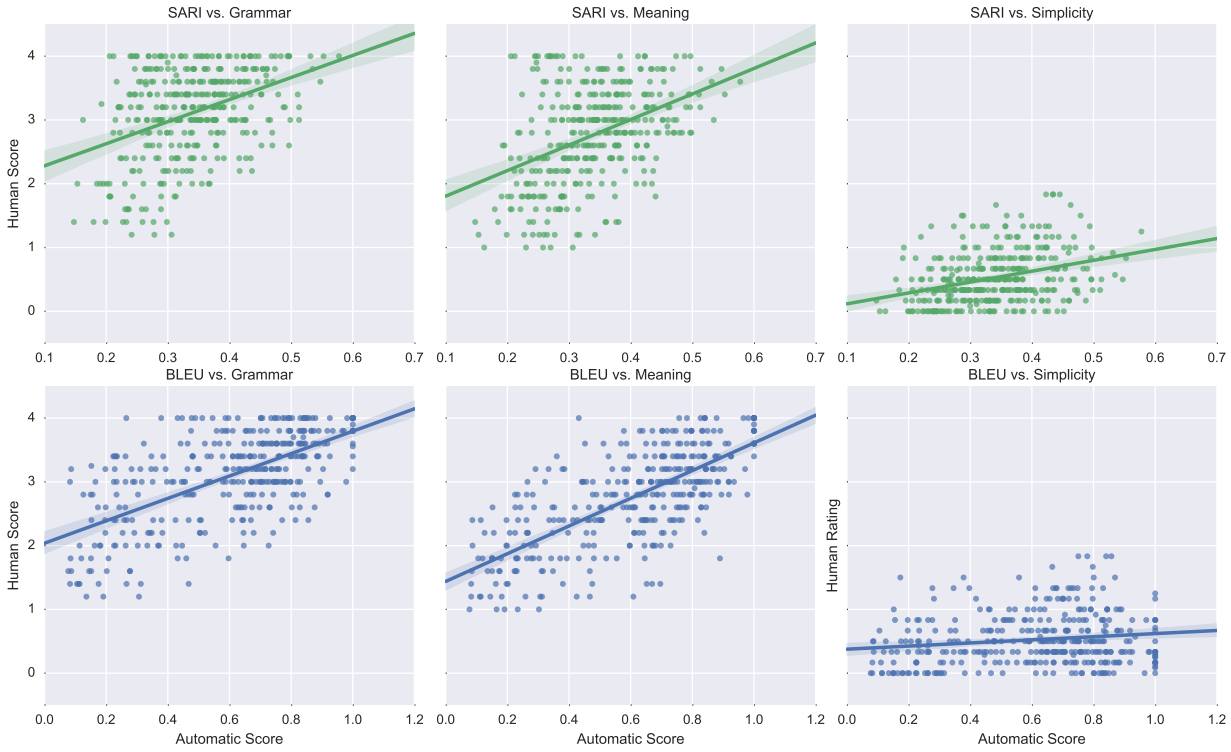

Figure 3: Scatter plots of automatic metrics against human scores for individual sentences.

works for multiple text-to-text generation tasks (including sentence simplification, compression and error correction), at the same time using the same idea of comparing system output against multiple references and against the input. The metric could possibly include tunable parameters or weighted human judgements on references to accommodate different tasks. Finally, we are also interested in designing neural translation models (Bahdanau et al., 2014; Sutskever et al., 2014; Chitnis and DeNero, 2015) for the simplification task.

Acknowledgments

The authors would like to thank Juri Ganitkevitch, Jonny Weese, Matt Post, Shashi Narayan, and Kristina Toutanova for valuable discussions. We also thank action editor Stefan Riezler and three anonymous reviewers for their thoughtful comments.

This material is based on research sponsored by the NSF under grant IIS-1430651. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of the NSF or the U.S. Government.

References

Allen, D. (2009). A study of the role of relative clauses in the simplification of news texts for learners of English. *System*, 37(4):585–599.

Amancio, M. A. and Specia, L. (2014). An analysis of crowdsourced text simplifications. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*.

Angrosh, M., Nomoto, T., and Siddharthan, A. (2014). Lexico-syntactic text simplification and compression with typed dependencies. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bannard, C. and Callison-Burch, C. (2005). Para-

phrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Biran, O., Brody, S., and Elhadad, N. (2011). Putting it simply: A context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*.

Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., and Tait, J. (1999). Simplifying text for language-impaired readers. In *Proceedings of the 14th Conference of the 9th European Conference for Computational Linguistics (EACL)*.

Chan, T. P., Callison-Burch, C., and Van Durme, B. (2011). Reranking bilingually extracted paraphrases using monolingual distributional similarity. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics (MTTG)*.

Chandrasekar, R., Doran, C., and Srinivas, B. (1996). Motivations and methods for text simplification. In *Proceedings of the 16th Conference on Computational linguistics (COLING)*.

Chen, B., Kuhn, R., and Larkin, S. (2012a). Port: A precision-order-recall MT evaluation metric for tuning. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chen, D. L. and Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chen, H.-B., Huang, H.-H., Chen, H.-H., and Tan, C.-T. (2012b). A simplification-translation-restoration framework for cross-domain SMT applications. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*.

Chitnis, R. and DeNero, J. (2015). Variable-length word encodings for neural translation models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Clarke, J. and Lapata, M. (2006). Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL-COLING)*.

Coster, W. and Kauchak, D. (2011a). Learning to simplify sentences using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*.

Coster, W. and Kauchak, D. (2011b). Simple English Wikipedia A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*.

Dahlmeier, D. and Ng, H. T. (2012). Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

De Belder, J. and Moens, M.-F. (2010). Text simplification for children. In *Proceedings of the SIGIR Workshop on Accessible Search Systems*.

Devlin, S., Tail, J., Canning, Y., Carroll, J., Minnen, G., and Pearce, D. (1999). The application of assistive technology in facilitating the comprehension of newspaper text by aphasic people. *Assistive Technology on the Threshold of the New Millennium*, page 160.

Evans, R., Orasan, C., and Dornescu, I. (2014). An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*.

Felice, M. and Briscoe, T. (2015). Towards a standard evaluation method for grammatical error detection and correction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Feng, L. (2008). Text simplification: A survey. *The City University of New York, Technical Report*.

Filippova, K. and Strube, M. (2008). Dependency tree based sentence compression. In *Proceedings*

of the Fifth International Natural Language Generation Conference (INLG).

Galley, M., Brockett, C., Sordoni, A., Ji, Y., Auli, M., Quirk, C., Mitchell, M., Gao, J., and Dolan, B. (2015). deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ganitkevitch, J., Callison-Burch, C., Napoles, C., and Van Durme, B. (2011). Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ganitkevitch, J., Cao, Y., Weese, J., Post, M., and Callison-Burch, C. (2012a). Joshua 4.0: Packing, PRO, and paraphrases. In *Proceedings of the Seventh Workshop on Statistical Machine Translation (WMT)*.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2012b). Monolingual distributional similarity for text-to-text generation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*.

Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2013). PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Gao, M., Xu, W., and Callison-Burch, C. (2015). Cost optimization in crowdsourcing translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Hopkins, M. and May, J. (2011). Tuning as ranking. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Horn, C., Manduca, C., and Kauchak, D. (2014). Learning a lexical simplifier using Wikipedia. In *Proceedings of the 52th Annual Meeting of the Associatioin for Computational Linguistics (ACL)*.

Hwang, W., Hajishirzi, H., Ostendorf, M., and Wu, W. (2015). Aligning sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Inui, K., Fujita, A., Takahashi, T., Iida, R., and Iwakura, T. (2003). Text simplification for reading assistance: A project note. In *Proceedings of the 2nd International Workshop on Paraphrasing (IWP)*.

Kaji, N., Kawahara, D., Kurohash, S., and Sato, S. (2002). Verb paraphrase based on case frame alignment. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*.

Kauchak, D. (2013). Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 2013 Conference of the Association for Computational Linguistics (ACL)*.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel. Technical report, Defence Technical Information Center (DTIC) Document.

Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*.

Liu, C., Dahlmeier, D., and Ng, H. T. (2010). TESLA: Translation evaluation of sentences with linear-programming-based analysis. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics (MATR)*.

Miwa, M., Saetre, R., Miyao, Y., and Tsujii, J. (2010). Entity-focused sentence simplification for relation extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*.

Napoles, C., Sakaguchi, K., Post, M., and Tetreault, J. (2015). Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Narayan, S. and Gardent, C. (2014). Hybrid simplification using deep semantics and machine trans-

lation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL)*.

Pavlick, E., Bos, J., Nissim, M., Beller, C., Durme, B. V., and Callison-Burch, C. (2015). Adding semantics to data-driven paraphrasing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Pavlick, E. and Nenkova, A. (2015). Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Pellow, D. and Eskenazi, M. (2014a). An open corpus of everyday documents for simplification tasks. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*.

Pellow, D. and Eskenazi, M. (2014b). Tracking human process using crowd collaboration to enrich data. In *Proceedings of Second AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Petersen, S. E. and Ostendorf, M. (2007). Text simplification for language learners: A corpus analysis. In *Proceedings of Workshop on Speech and Language Technology for (SLaTE)*.

Post, M., Ganitkevitch, J., Orland, L., Weese, J., Cao, Y., and Callison-Burch, C. (2013). Joshua 5.0: Sparser, better, faster, server. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT)*.

Rello, L., Baeza-Yates, R. A., and Saggion, H. (2013). The impact of lexical simplification by verbal paraphrases for people with and without dyslexia. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*.

Siddharthan, A. (2006). Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Siddharthan, A. (2014). A survey of research on text simplification. *Special issue of International Journal of Applied Linguistics*, 165(2).

Siddharthan, A. and Katsos, N. (2010). Reformulating discourse connectives for non-expert readers. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Siddharthan, A., Nenkova, A., and McKeown, K. (2004). Syntactic simplification for improving content selection in multi-document summarization. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*.

Specia, L., Jauhar, S. K., and Mihalcea, R. (2012). SemEval-2012 task 1: English lexical simplification. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval)*.

Štajner, S., Béchara, H., and Saggion, H. (2015). A deeper exploration of the standard PB-SMT approach to text simplification and its evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Štajner, S., Mitkov, R., and Saggion, H. (2014). One step closer to automatic evaluation of text simplification systems. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*.

Sun, H. and Zhou, M. (2012). Joint learning of a dual SMT system for paraphrase generation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*.

Vickrey, D. and Koller, D. (2008). Sentence simplication for semantic role labeling. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*.

Watanabe, W. M., Junior, A. C., Uzêda, V. R., Fortes, R. P. d. M., Pardo, T. A. S., and Aluísio, S. M. (2009). Facilita: Reading assistance

for low-literacy readers. In *Proceedings of the 27th ACM International Conference on Design of Communication (SIGDOC)*.

Woodsend, K. and Lapata, M. (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)*.

Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Xu, W., Callison-Burch, C., and Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics (TACL)*, 3:283–297.

Xu, W., Ritter, A., Dolan, B., Grishman, R., and Cherry, C. (2012). Paraphrasing for style. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*.

Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., and Lee, L. (2010). For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*.