A Deeper Exploration of the Standard PB-SMT Approach to Text Simplification and its Evaluation

Sanja Štajner[1] and **Hannah Béchara**[1] and **Horacio Saggion**[2]
[1]Research Group in Computational Linguistics, University of Wolverhampton, UK
[2]TALN Research Group, Universitat Pompeu Fabra, Spain
{SanjaStajner,Hanna.Bechara}@wlv.ac.uk, horacio.saggion@upf.edu

Abstract

In the last few years, there has been a growing number of studies addressing the Text Simplification (TS) task as a monolingual machine translation (MT) problem which translates from 'original' to 'simple' language. Motivated by those results, we investigate the influence of quality vs quantity of the training data on the effectiveness of such a MT approach to text simplification. We conduct 40 experiments on the aligned sentences from English Wikipedia and Simple English Wikipedia, controlling for: (1) the similarity between the original and simplified sentences in the training and development datasets, and (2) the sizes of those datasets. The results suggest that in the standard PB-SMT approach to text simplification the quality of the datasets has a greater impact on the system performance. Additionally, we point out several important differences between cross-lingual MT and monolingual MT used in text simplification, and show that BLEU is not a good measure of system performance in text simplification task.

1 Introduction

In the last few years, a growing number of studies have addressed the text simplification (TS) task as a monolingual machine translation (MT) problem of translating sentences from 'original' to 'simple' language. Several studies reported promising results using standard phrase-based statistical machine translation (PB-SMT) for this task (Specia, 2010; Coster and Kauchak, 2011a; Wubben et al., 2012), but made no attempt to explain the reasons behind the success of their systems. Specia (2010) obtained reasonably good results (BLEU = 60.75) despite the small size of the datasets used (4,483 original sentences and their corresponding simplifications). Her results indicated that in this specific monolingual MT task, we do not need such large datasets (as in cross-lingual MT) in order to achieve good results.

At the moment, the scarcity and very limited sizes of the available TS datasets (usually only up to 1,000 sentence pairs) are the main factors which impede the use of data-driven approaches to text simplification for all languages except English (for which English Wikipedia and Simple English Wikipedia offer a large comparable TS dataset). Therefore, in this paper, we decided to investigate several important issues in MT-based text simplification:

1. The impact of the size of the training and development datasets;

2. The impact of the similarity between the original and simplified sentences in the training and development datasets; and

3. The suitability of using the BLEU score for the automatic evaluation of system's performance.

To the best of our knowledge, there have been no studies which address those important questions.

In order to explore the first two issues, we conduct 40 translation experiments using the aligned sentence pairs from the largest existing TS corpus (Wikipedia TS corpus), controlling the training and development datasets for: (1) sentence similarity (in terms of the S-BLEU score), and (2) size. Our results indicate that only the former can influence the MT output significantly. In order to explore the last issue, we test our models on two different test sets and perform human evaluation of the output of several systems.

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 823–828,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

2 Related Work

Specia (2010) used the standard PB-SMT model provided by the Moses toolkit (Koehn et al., 2007) to translate from 'original' to 'simple' sentences in Brazilian Portuguese. The dataset contained manual simplifications aimed at people with low literacy levels. The most commonly used simplifications (by human editors) were lexical substitutions and splitting sentences (Gasperin et al., 2009). In terms of the automatic BLEU evaluation (Papineni et al., 2002), the results were reasonably good (BLEU = 60.75) despite the small size of the corpora (4,483 original sentences and their corresponding simplifications). However, the TS system was overcautious in performing simplifications, i.e. the simplifications produced by the systems were closer to the source than to the reference segments (Specia, 2010).

Coster and Kauchak (2011a) used the same approach for English. Additionally, they extended the PB-SMT system by adding phrasal deletion to the probabilistic translation model in order to better cover deletion, which is a frequent phenomenon in TS. The system was trained on 124,000 aligned sentences from English Wikipedia and Simple English Wikipedia. The analysis of the Wikipedia TS corpus (Coster and Kauchak, 2011b) reported that rewordings (1–1 lexical substitutions) are the most common simplification operation (65%). The system with added phrasal deletion achieved the BLEU score of 60.46, while the the standard model without phrasal deletion achieved the BLEU score of 59.87. However, the baseline (BLEU score when the system does not perform any simplification on the original sentence) was 59.37, indicating that the systems often leave the original sentences unchanged. In order to address that problem, Wubben et al. (2012) performed post-hoc reranking on the Moses' output (simplification hypotheses) based on their dissimilarity to the input (original sentences), while at the same time controlling for its adequacy and fluency.

Štajner (2014) applied the same PB-SMT model to two different TS corpora in Spanish, which contained different levels of simplification. The results, which should be regarded only as preliminary as both corpora have fewer than 1,000 sentence pairs, imply that the level of simplification in the training datasets has a greater impact than the size of the datasets on the system's performance.

3 Methodology

We focus on the two TS corpora available for English (Wikipedia and EncBrit) and train a series of translation models on training and development datasets of varying size and quality.

3.1 Corpora

Wikipedia is a comparable TS corpus of 137,000 automatically aligned sentence pairs from English Wikipedia and Simple English Wikipedia[1], previously used by Coster and Kauchak (2011a). We use a small portion of this corpus (240 sentence pairs) to build the first test set (WikiTest), and 88,000 sentence pairs from the remaining sentence pairs to build translation models.

EncBrit is a comparable TS corpus of original sentences from Encyclopedia Britannica and their manually simplified versions for children (Barzilay and Elhadad, 2003).[2] Given its small size (601 sentence pairs) this dataset is not used in the translation experiments. It is only used as the second test set (EncBritTest).

3.2 Experimental Setup

In all experiments, we use the same standard PB-SMT model (Koehn et al., 2007), the GIZA++ implementation of IBM word alignment model 4 (Och and Ney, 2003), and the refinement and phrase-extraction heuristics described further by Koehn *et al.* (2003). We tune the systems using minimum error rate training (MERT) (Och, 2003). For the language model (LM) we use the corpus of 60,000 Simple English Wikipedia articles[3] and build a 3-gram language model with Kneser-Ney smoothing trained with SRILM (Stolcke, 2002). We limit our stack size to 500 hypotheses during decoding.

3.3 Training and development datasets

We tokenise and shuffle the initial dataset of 167,689 aligned sentences from the Wikipedia dataset.[4] Using the simplified sentences as references and the original sentences as hypotheses,

[1] http://www.cs.middlebury.edu/~dkauchak/simplification/

[2] http://www.cs.columbia.edu/~noemie/alignment/

[3] Version 2.0 document-aligned data, available at: http://www.cs.middlebury.edu/~dkauchak/simplification/

[4] Version 2.0 sentence-aligned data, available at: http://www.cs.middlebury.edu/~dkauchak/simplification/

Table 1: Examples of sentences pairs with various S-BLEU scores from the training sets

S-BLEU	Original sentence	Simpler version
0.08	*In women, the larger* mammary glands *within* the breast *produce the milk*.	The breast *contains* mammary glands.
0.38	*Built as a double-track railroad bridge, it* was completed on January 1, 1889, and *went out of service* on May 8, 1974.	*It was built for trains and* was completed on January 1, 1889. *It closed down* on May 8, 1974 *after a bad fire*.
0.55	In 2000, the series *sold its naming rights to* Internet search engine Northern Light *for five seasons, and t*he series was named the Indy Racing Northern Light Series.	In 2000, the series *sponsor became the* Internet search engine Northern Light. *T*he series was named the Indy Racing Northern Light Series.
0.63	*Wildlife which eat acorns as* an important part of their diet*s* include birds, such as jays, pigeons, some ducks, and several species of woodpeckers.	*Creatures that make acorns* an important part of their diet include birds, such as jays, pigeons, some ducks and several species of woodpeckers.
0.77	It was *discovered* by Brett J. Gladman in 2000, and given the *temporary* designation S2000 S 5.	It was *found* by Brett J. Gladman in 2000, and given the designation S2000 S 5.
0.87	Austen was not well known in Russia *and t*he first Russian translation of an Austen novel did not appear until 1967.	Austen was not well known in Russia. *T*he first Russian translation of an Austen novel did not appear until 1967.

we rank each sentence pair by its sentence-wise BLEU (S-BLEU) score and categorise the sentence pairs into eight different sets depending on the interval in which their S-BLEU scores lie ((0, 0.3], (0.3, 0.4], (0.4, 0.5], (0.5, 0.6], (0.6, 0.7], (0.7, 0.8], (0.8, 0.9], (0.9, 1]). With each of the eight sets, we train five translation models, varying the number of sentences used for training and tuning (2,000, 4,000, 6,000, 8,000, and 10,000 for training and 200, 400, 600, 800, and 1,000 for tuning, respectively). That leads to a total of 40 translation models varying by number of sentence pairs and similarity between original and simplified sentences (in terms of the S-BLEU score) in the datasets used for their training and tuning. Several examples of sentence pairs with various S-BLEU scores are presented in Table 1.

3.4 Test datasets

We test our models on two different test sets:

1. The **WikiTest** which contains a total of 240 sentence pairs, with 30 sentence pairs from each of the eight categories with different intervals for the S-BLEU scores ([0,0.3], (0.3,0.4], ... , (0.9,1]);

2. The **EncBritTest** which contains all 601 sentence pairs present in the EncBrit corpus (with an unbalanced number of sentence pairs from each of the eight S-BLEU intervals).

The sizes of both test sets and their BLEU scores (calculated using the original sentences as simplification/translation hypotheses and the corresponding manually simplified sentences as simplification/translation references) are given in Table 2. Note that those BLEU scores can be regarded as the baselines for the translation experiments, as they correspond to the BLEU score obtained when the systems do not perform any changes to the input.

Table 2: Test sets for all translation experiments

Test set	Size	BLEU
WikiTest	240	62.27
EncBritTest	601	12.40

4 Automatic Evaluation

The BLEU scores for all 40 experiments tested on the WikiTest dataset, are presented in Table 3. The baseline BLEU score (when no simplification is performed) for this test set is 62.27 (Table 2). As shown in Table 3, none of the 40 experiments have even reached that baseline. We compare S-BLEU scores for each pair of experiments (240 reference sentences in the test set and their corresponding automatically simplified sentences) using the paired t-test in SPSS in order to check whether the differences in the obtained results are significant. The only results that are significantly lower than the rest are those obtained for the experiments in which the training and development datasets consist only of the sentence pairs with S-BLEU scores between 0 and 0.3. The results sug-

Table 3: BLEU scores on the WikiTest dataset

S-BLEU	Size of the training set				
	2,000	4,000	6,000	8,000	10,000
[0, 0.3]	56.38	56.38	56.15	57.75	57.89
(0.3, 0.4]	60.89	61.35	61.76	61.52	61.37
(0.4, 0.5]	61.27	61.36	61.74	61.55	**62.11**
(0.5, 0.6]	60.96	61.30	61.52	61.77	61.98
(0.6, 0.7]	60.96	61.30	61.60	61.69	61.80
(0.7, 0.8]	61.56	61.38	61.67	61.77	61.89
(0.8, 0.9]	61.54	61.49	61.51	61.57	61.61
(0.9, 1]	61.57	61.57	61.59	61.55	61.55

The rows represent intervals of the S-BLEU scores on the training and development datasets, while the columns represent the number of the sentence pairs used for training. The highest score is presented in bold; the baseline (no simplification performed) is 62.27.

Table 4: BLEU scores on the EncBritTest dataset

S-BLEU	Size of the training set				
	2,000	4,000	6,000	8,000	10,000
[0, 0.3]	13.84	13.84	13.87	13.68	13.59
(0.3, 0.4]	14.05	13.95	14.08	14.06	14.01
(0.4, 0.5]	14.02	14.09	14.17	14.15	14.12
(0.5, 0.6]	14.09	14.22	14.27	14.16	14.13
(0.6, 0.7]	14.25	14.30	14.35	14.35	14.32
(0.7, 0.8]	14.30	14.29	14.30	14.30	14.28
(0.8, 0.9]	14.38	14.40	14.40	14.40	**14.41**
(0.9, 1]	12.71	12.52	12.46	12.39	12.54

The rows represent intervals of the S-BLEU scores on the training and development datasets, while the columns represent the number of the sentence pairs used for training. The highest score is presented in bold; the baseline (no simplification performed) is 12.40.

Table 5: Systems used in human evaluation

System	Training size	Dev. size	S-BLEU
S-03-200	2,000	200	[0,0.3]
S-03-1000	10,000	1,000	[0,0.3]
S-06-200	2,000	200	(0.5,0.6]
S-06-1000	10,000	1,000	(0.5,0.6]
S-10-200	2,000	200	[0.9,1]
S-10-1000	10,000	1,000	[0.9,1]

gest that the sizes of the training and development datasets do not influence the translation results significantly on any type of sentence pairs used.

The results of the experiments tested on EncBritTest (Table 4) again show that the quantity of the training data does not influence system performance. There are no statistically significant differences (measured by the paired t-test on S-BLEU scores on all 601 reference sentences and the corresponding automatic simplifications) among experiments which differ only in the size of the training and development datasets. However, the models trained and tuned on the datasets consisting of the sentence pairs with the highest and the lowest S-BLEU scores ([0,0.3] and (0.9,1]) perform significantly worse than the models trained and tuned on the sentence pairs with S-BLEU scores belonging to other intervals.

5 Human Evaluation

The results presented in Tables 3 and 4 indicate that the BLEU score, in MT-based text simplification, mostly reflects the surface similarity of the original and simplified sentences in the test set and does not give an informative evaluation of the systems. Therefore, we conducted a human assessment of the generated sentences. Following the standard procedure for human evaluation of TS systems used in previous studies (Coster and Kauchak, 2011a; Drndarević et al., 2013; Wubben et al., 2012; Saggion et al., 2015), three human evaluators were asked to assess the generated sentences on a 1–5 scale (where the higher mark always denotes better output) according to three criteria: grammaticality (G), meaning preservation (M), and simplicity (S).

We decided that the same person has to rate all simplified versions of the same original sentence (shown always in a random order), in order to make a fairer comparison among the systems. That decision, however, limited the number of systems we can evaluate. Therefore, we focused only on six out of 40 trained systems (Table 5). Several examples of the automatically simplified sentences and their scores are presented in Table 6.

The results of the human evaluation are given in Table 7. It seems that the use of the sentence pairs with the S-BLEU score between 0.5 and 0.6 leads to the best system performances in terms of grammaticality and meaning preservation, while at the same time improving the simplicity of the sentences.[5] Furthermore, the differences in human scores between the systems differing only in size of the datasets used were not statistically significant. At the same time, the differences in human

[5] The details of the human evaluation and examples can be found in (Štajner, 2015).

Table 6: Outputs of different systems and their human evaluation scores

System	Sentence	G	M	S
Original	Madrid was occupied by French troops during the Napoleonic Wars, and Napoleon's brother Joseph was installed on the throne.	5	/	4
S-03-200	Madrid was occupied by French *his soldiers* during the Napoleonic Wars, and Napoleon's brother Joseph was installed on the throne.	4	4	4
S-03-1000	Madrid was occupied by French troops during the Napoleonic Wars, and Napoleon's brother Joseph was *put* on the throne.	5	5	5
S-10-1000	Madrid was occupied by French troops during the Napoleonic Wars, and Napoleon's brother Joseph was *-RRB-* installed *on them* on the throne.	3	3	3
Original	Although largely of postwar construction, this central area retains its old street pattern, and most of the surviving historical and architectural monuments are located there.	5	/	2
S-03-200/1000	Although largely of postwar construction, this central area retains its old street pattern, and most of the surviving *and architectural historical* monuments are located there.	5	4	3
S-06-1000	Although *mostly* of postwar construction, this central area retains its old street pattern, and most of the surviving historical and architectural monuments are located there.	5	5	2
S-10-200	Although largely of postwar construction, this central area retains its old street pattern, and most of the surviving historical monuments and architectural *are a instead*.	3	3	2
S-10-1000	*As* of *the* postwar construction, *in* this central area *uses* its old street pattern, and most of the *historical monuments and and architectural* are located there.	2	3	2

The columns *G*, *M*, *S* contain the mean value of the human scores for grammaticality, meaning preservation, and simplicity, respectively. Differences to the original versions are shown in italics. Systems which are not presented did not make any changes to these two original sentences.

Table 7: Results of the human evaluation

System	G	M	S
Original	4.85	/	2.60
S-03-200	4.03	3.95	2.57
S-03-1000	4.20	4.03	**2.85**
S-06-200	**4.50**	4.45	2.68
S-06-1000	4.43	**4.48**	2.72
S-10-200	3.25	2.92	2.45
S-10-1000	2.92	2.95	2.53

The mean value of the human scores for grammaticality (*G*), meaning preservation (*M*), and simplicity (*S*). The highest achieved scores (excluding the scores for original sentences) on each aspect (G, M, and S) are presented in bold.

scores between the systems differing only in similarity of the sentence pairs (the interval of the S-BLEU score) used were statistically significant.

6 Conclusions

Recently, there have been several attempts at addressing the TS task as a monolingual translation problem, translating from 'original' to 'simple' sentences. However, they did not try to seek reasons for the success or the failure of their systems.

Our experiments, conducted on 40 different, carefully designed datasets from the largest available sentence-aligned TS corpus (Wikipedia TS corpus), provide valuable insights into how much of an effect the size and the quality of the training data have on the performance of the PB-SMT system which tries to learn to translate from 'original' to 'simple' sentences. The results indicate that using the sentence pairs with low S-BLEU scores for training and tuning of PB-SMT models for TS tend to cause the fluency to deteriorate and even change the meaning of the output. Furthermore, it seems that the sizes of the training and development datasets do not play a significant role in how successful the model is. It appears that carefully selected sentence pairs in the training and development datasets (i.e. sentence pairs with a moderate similarity) lead to best performances of PB-SMT systems regardless of the size of the datasets.

Our results open up new directions for enhancing the current PB-SMT models for TS, indicating that their performance can be significantly improved by carefully filtering sentence pairs used for training and tuning.

Acknowledgements

The research described in this paper was partially funded by the project SKATER-UPF-TALN (TIN2012-38584-C06-03), Ministerio de Economía y Competitividad, Secretaría de Estado de Investigación, Desarrollo e Innovación, Spain, and the project ABLE-TO-INCLUDE (CIP-ICT-PSP-2013-7/621055). Hannah Béchara is supported by the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 31747.

References

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 25–32. Association for Computational Linguistics.

William Coster and David Kauchak. 2011a. Learning to Simplify Sentences Using Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–9. Association for Computational Linguistics.

William Coster and David Kauchak. 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)*, pages 665–669. Association for Computational Linguistics.

Biljana Drndarević, Sanja Štajner, Stefan Bott, Susana Bautista, and Horacio Saggion. 2013. Automatic Text Simplication in Spanish: A Comparative Evaluation of Complementing Components. In *Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, volume 7817 of *Lecture Notes in Computer Science*, pages 488–500. Springer Berlin Heidelberg.

Caroline Gasperin, Lucia Specia, Tiago F. Pereira, and Sandra M. Aluísio. 2009. Learning When to Simplify Sentences for Natural Text Simplification. In *Proceedings of the Encontro Nacional de Intelignncia Artificial (ENIA), Bento Gonalves, Brazil*, pages 809–818.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 48–54. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

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

Franz Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167. Association for Computational Linguistics.

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

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Transactions on Accessible Computing*, 6(4):14:1–14:36.

Lucia Specia. 2010. Translating from complex to simplified sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language (PROPOR)*, volume 6001 of *Lecture Notes in Computer Science*, pages 30–39. Springer Berlin Heidelberg.

Andreas Stolcke. 2002. SRILM - an Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 901–904.

Sanja Štajner. 2014. Translating sentences from 'original' to 'simplified' spanish. *Procesamiento del Lenguaje Natural*, 53:61–68.

Sanja Štajner. 2015. *New Data-Driven Approaches to Text Simplification*. Ph.D. thesis, University of Wolverhampton, UK.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL): Long Papers - Volume 1*, pages 1015–1024. Association for Computational Linguistics.