Sentence Alignment Methods for Improving Text Simplification Systems

Sanja Štajner[1], **Marc Franco-Salvador**[2,3],
Simone Paolo Ponzetto[1], **Paolo Rosso**[3], **Heiner Stuckenschmidt**[1]

[1] DWS Research Group, University of Mannheim, Germany
[2] Symanto Research, Nuremberg, Germany
[3] PRHLT Research Center, Universitat Politècnica de València, Spain
{sanja,simone,heiner}@informatik.uni-mannheim.de
marc.franco@symanto.net, prosso@prhlt.upv.es

Abstract

We provide several methods for sentence-alignment of texts with different complexity levels. Using the best of them, we sentence-align the Newsela corpora, thus providing large training materials for automatic text simplification (ATS) systems. We show that using this dataset, even the standard phrase-based statistical machine translation models for ATS can outperform the state-of-the-art ATS systems.

1 Introduction

Automated text simplification (ATS) tries to automatically transform (syntactically, lexically and/or semantically) complex sentences into their simpler variants without significantly altering the original meaning. It has attracted much attention recently as it could make texts more accessible to wider audiences (Aluísio and Gasperin, 2010; Saggion et al., 2015), and used as a pre-processing step, improve performances of various NLP tasks and systems (Vickrey and Koller, 2008; Evans, 2011; Štajner and Popović, 2016).

However, the state-of-the-art ATS systems still do not reach satisfying performances and require some human post-editing (Štajner and Popović, 2016). While the best supervised approaches generally lead to grammatical output with preserved original meaning, they are overcautious, making almost no changes to the input sentences (Specia, 2010; Štajner et al., 2015), probably due to the limited size or bad quality of parallel TS corpora used for training. The largest existing sentence-aligned TS dataset for English is the English Wikipedia – Simple English Wikipedia (EW–SEW) dataset, which contains 160-280,000 sentence pairs, depending on whether we want to model only traditional sentence rewritings or also to model content reduction and stronger paraphrasing (Hwang et al., 2015). For Spanish, the largest existing parallel TS corpus contains only 1,000 sentence pairs thus impeding the use of fully supervised approaches. The best unsupervised lexical simplification (LS) systems for English which leverage word-embeddings (Glavaš and Štajner, 2015; Paetzold and Specia, 2016) seem to perform more lexical substitutions but at the cost of having less grammatical output and more often changed meaning. However, there have been no direct comparisons of supervised and unsupervised state-of-the-art approaches so far.

The Newsela corpora[1] offers over 2,000 original news articles in English and around 250 in Spanish, manually simplified to 3–4 different complexity levels following strict guidelines (Xu et al., 2015). Although it was suggested that it has better quality than the EW–SEW corpus (Xu et al., 2015), Newsela has not yet been used for training end-to-end ATS systems, due to the lack of its sentence (and paragraph) alignments. Such alignments, between various text complexity levels, would offer large training datasets for modelling different levels of simplification, i.e. 'mild' simplifications (using the alignments from the neighbouring levels) and 'heavy' simplifications (using the alignments of level pairs: 0–3, 0–4, 1–4).

Contributions. We: (1) provide several methods for paragraph- and sentence alignment of parallel texts, and for assessing similarity level between pairs of text snippets, as freely avail-

[1] Freely available: https://newsela.com/data/

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 97–102
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-2016

able software;[2] (2) compare the performances of lexically- and semantically-based alignment methods across various text complexity levels; (3) test the hypothesis that the original order of information is preserved during manual simplification (Bott and Saggion, 2011) by offering customized MST-LIS alignment strategy (Section 3.1); and (4) show that the new sentence-alignments lead to the state-of-the-art ATS systems even in a basic phrase-based statistical machine translation (PB-SMT) approach to text simplifications.

2 Related Work

The current state-of-the-art systems for automatic sentence-alignment of original and manually simplified texts are the GSWN method (Hwang et al., 2015) used for sentence-alignment of original and simple English Wikipedia, and the HMM-based method (Bott and Saggion, 2011) used for sentence-alignment of the Spanish Simplext corpus (Saggion et al., 2015).

The HMM-based method can be applied to any language as it does not require any language-specific resources. It is based on two hypotheses: (H1) that the original order of information is preserved, and (H2) that every 'simple' sentence has at least one corresponding 'original' sentence (it can have more than one in the case of 'n-1' or 'n-m' alignments).

As Simple Wikipedia does not represent direct simplification of the 'original' Wikipedia articles ('simple' articles were written independently of the 'original' ones), GSWN method does not assume H1 or H2. The main limitations of this method are that it only allows for '1-1' sentence alignments – which is very restricting for TS as it does not allow for sentence splitting ('1-n'), and summarisation and compression ('n-1' and 'n-m') alignments – and it is language-dependent as it requires English Wiktionary.

Unlike the GSWN method, all the methods we apply are language-independent, resource-light and allow for '1-n', 'n-1', and 'n-m' alignments. Similar to the HMM-method, our methods assume the hypothesis H2. We provide them in both variants, using the hypothesis H1 and without it (Section 3.1).

3 Approach

Having a set of 'simple' text snippets S and a set of 'complex' text snippets C, we offer two strategies (Section 3.1) to obtain the alignments (s_i, c_j), where $s_i \in S$, $c_j \in C$. Each alignment strategy, in turn, can use one of the three methods (Section 3.2) to calculate similarity scores between text snippets (either paragraphs or sentences).

3.1 Alignment strategies

Most Similar Text (MST): Given one of the similarity methods (Section 3.2), MST compares similarity scores of all possible pairs (s_i, c_j), and aligns each $s_i \in S$ with the closest one in C.

MST with Longest Increasing Sequence (MST-LIS): MST-LIS uses the hypothesis H1. It first uses the MST strategy, and then postprocess the output by extracting – from all obtained alignments – only those alignments $l_i \in L$, which contain the longest increasing sequence of offsets j_k in C. In order to allow for '1–n' alignments (i.e. sentence splitting), we allow for repeated offsets of C ('complex' text snippets) in L. The 'simple' text snippets not contained in L are included in the set U of unaligned snippets. Finally, we align each $u_m \in U$ by restricting the search space in C to those offsets of 'complex' text snippets that correspond to the previous and the next aligned 'simple' snippets. For instance, if $L = \{(s_1, c_4), (s_3, c_7)\}$ and $U = \{s_2\}$, then the search space for the alignments of s_2 is reduced to $\{c_4...c_7\}$. We denote this strategy with an '*' in the results (Table 2), e.g. C3G*.

3.2 Similarity Methods

C3G: We employ the Character N-Gram (CNG) (Mcnamee and Mayfield, 2004) similarity model (for $n = 3$) with log TF-IDF weighting (Salton and McGill, 1986) and compare vectors using the cosine similarity.

WAVG: We use the continuous skip-gram model (Mikolov et al., 2013b) of the TensorFlow toolkit[3] to process the whole English Wikipedia and generate continuous representations of its words.[4] For each text snippet, we average its word vectors to obtain a single representation of its content as this setting has shown good results

[2] https://github.com/neosyon/SimpTextAlign

[3] https://www.tensorflow.org/

[4] We use 300-dimensional vectors, context windows of size 10, and 20 negative words for each sample, in all our continuous word-based models.

Match	Transformation	Original	Simple
Full	syntactic simplification; reordering of sentence constituents	During the 13th century, gingerbread was brought to Sweden by German immigrants.	German immigrants brought it to Sweden during the 13th century.
Full	lexical paraphrasing	During the 13th century, gingerbread was brought to Sweden by German immigrants.	German immigrants brought it to Sweden during the 13th century.
Partial	strong paraphrasing	Gingerbread foods vary, ranging from a soft, moist loaf cake to something close to a ginger biscuit.	Gingerbread is a word which describes different sweet food products from soft cakes to a ginger biscuit.
Partial	adding explanations	Humidity is the amount of water vapor in the air.	Humidity (adjective: humid) refers to water vapor in the air, but not to liquid droplets in fog, clouds, or rain.
Partial	sentence compression	Falaj irrigation is an ancient system dating back thousands of years and is used widely in Oman, the UAE, China, Iran and other countries.	The ancient falaj system of irrigation is still in use in some areas.

Table 1: Examples of full and partial matches from the EW–SEW dataset (Hwang et al., 2015).

in other NLP tasks (e.g. for selecting out-of-the-list words (Mikolov et al., 2013a)). Finally, the similarity between text snippets is estimated using the cosine similarity.

CWASA: We employ the Continuous Word Alignment-based Similarity Analysis (CWASA) model (Franco-Salvador et al., 2016), which finds the optimal word alignment by computing cosine similarity between continuous representations of all words (instead of averaging word vectors as in the case of WAVG). It was originally proposed for plagiarism detection with excellent results, especially for longer text snippets.

4 Manual Evaluation

To compare the performances of different alignment methods, we randomly selected 10 original texts (Level 0) and their corresponding simpler versions at Levels 1, 3 and 4. Instead of creating a 'gold standard' and then automatically evaluating the performances, we asked two annotators to rate each pair of automatically aligned paragraphs and sentences – by each of the possible six alignment methods and the HMM-based method (Bott and Saggion, 2011) – for three pairs of text complexity levels (0–1, 0–4, and 3–4) on a 0–2 scale, where: **0** – no semantic overlap in the content; **1** – partial semantic overlap (*partial matches*); **2** – same semantic content (*good matches*). This resulted in a total of 1526 paragraph- and 1086 sentence-alignments for the 0–1 pairs, and 1218 paragraph- and 1266 sentence-alignments for the 0–4 and 3–4 pairs. In the context of TS, both good- and partial matches are important. While full semantic overlap models full paraphrases ('1-1' alignments), partial overlap models sentence splitting ("1-n" alignments), deleting irrelevant sentence parts, adding explanations, or summarizing ('n-m' alignments). Several examples of full and partial matches from the EW–SEW dataset (Hwang et al., 2015) are given in Table 1.

We expect that the automatic-alignment task is the easiest between the 0–1 text complexity levels, and much more difficult between the 0-4 levels (Level 4 is obtained after four stages of simplification and thus contains stronger paraphrases and less lexical overlap with Level 0 than Level 1 has). We also explore whether the task is equally difficult whenever we align two neighbouring levels, or the difficulty of the task depends on the level complexity (0–1 vs. 3–4). The obtained inter-annotator agreement, weighted Cohen's κ (on 400 double-annotated instances) was between 0.71 and 0.74 depending on the task and levels.

The results of the manual analysis (Table 2) showed that: (1) all applied methods significantly ($p < 0.001$) outperformed the HMM method on both paragraph- and sentence-alignment tasks;[5] (2) the methods which do not assume hypothesis H1 (C3G, CWASA, and WAVG) led to (not significantly) higher percentage of correct alignments than their counterparts which do assume

[5] Although some of our methods share the same percentage of good+partial matches with the HMM method on the paragraph-alignment 0–1 task, there is still significant difference in the obtained scores (in some cases, our methods led to good matches whereas the HMM only led to partial matches).

Method	Sentence			Paragraph		
	0–1	0–4	3–4	0–1	0–4	3–4
C3G	98.3	56.1	81.1	**98.6**	**86.8**	**95.2**
C3G*	96.7	54.7	78.8	95.4	77.0	92.3
CWASA	98.3	45.3	79.7	98.2	83.3	94.1
CWASA*	96.1	42.1	76.4	95.4	74.1	90.5
WAVG	97.8	56.1	79.7	96.8	75.9	91.7
WAVG*	96.1	50.0	79.7	96.8	69.5	89.3
C3G-2s	**98.5**	**57.8**	**83.5**	/	/	/
HMM	86.2	25.2	65.6	96.8	41.2	65.5

Table 2: Percentage of good+partial sentence- and paragraph-alignments on the English Newsela corpus. All results are significantly better ($p < 0.001$, Wilcoxon's signed rank test) than those obtained by the HMM method (Bott and Saggion, 2011). The best scores are in bold.

H1 (C3G*, CWASA*, WAVG*); (3) the difference in the performances of the lexical approach (C3G) and semantic approaches (CWASA and WAVG) was significant only in the 0–4 sentence-alignment task, where CWASA performed significantly worse ($p < 0.001$) than the other two methods, and in the 0–4 paragraph-alignment task, where WAVG performed significantly worse than C3G; (4) the 2-step C3G alignment-method (*C3G-2s*), which first aligns paragraphs using the best paragraph-alignment method (C3G) and then within each paragraph align sentences with the best sentence-alignment method (C3G), led to more good+partial alignments than the 'direct' sentence-alignment C3G method.

5 Extrinsic Evaluation

Finally, we test our new English Newsela (C3G-2s) sentence-alignments (both for the neighbouring levels – *neighb.* and for all levels – *all*) and Newsela sentence-alignments for neighboring levels obtained with HMM-method[6] (Bott and Saggion, 2011) in the ATS task using standard PB-SMT models[7] in the Moses toolkit (Koehn et al., 2007). We vary the training dataset and the corpus used to build language models (LMs), while keeping always the same 2,000 sentence pairs for tuning (Xu et al., 2016) and the first 70 sentence pairs of their test set[8] for our human evaluation. Using that particular test set allow us to compare our (PBSMT) systems with the output of the state-of-the-art syntax-based MT (SBMT) system for TS (Xu et al., 2016) which is not freely available. We compare: (1) the performance of the standard PBSMT model which uses only the already available EW–SEW dataset (Hwang et al., 2015) with the performances of the same PBSMT models but this time using the combination of the EW–SEW dataset and our newly-created Newsela datasets; (2) the latter PBSMT models (which use both EW–SEW and new Newsela datasets) against the state-of-the-art supervised ATS system (Xu et al., 2016), and one of the recently proposed unsupervised lexical simplification systems, the LightLS system (Glavaš and Štajner, 2015).[9]

We perform three types of human evaluation on the outputs of all systems. First, we count the total number of changes made by each system (*Total*), counting the change of a whole phrase (e.g. "*become defunct*" → "*was dissolved*") as one change. We mark as *Correct* those changes that preserve the original meaning and grammaticality of the sentence (assessed by two native English speakers) and, at the same time, make the sentence easier to understand (assessed by two non-native fluent English speakers).[10] Second, three native English speakers rate the grammaticality (*G*) and meaning preservation (*M*) of each sentence with at least one change on a 1–5 Likert scale (1 – very bad; 5 – very good). Third, the three non-native fluent English speakers were shown original (reference) sentences and target (output) sentences (one pair at the time) and asked whether the target sentence is: +2 – much simpler; +1 – somewhat simpler; 0 – equally difficult; -1 – somewhat more difficult; -2 – much more difficult, than the reference sentence. While the correctness of changes takes into account the influence of each individual change on grammaticality, meaning and simplicity of a sentence, the *Scores (G and M)* and *Rank (S)* take into account the mutual influence of all changes within a sentence.

Adding our sentence-aligned Newsela corpus

[6] Given that the performance of the HMM-method was poor for non-neighboring levels (Table 2).

[7] GIZA++ implementation of the IBM word alignment model 4 (Och and Ney, 2003), refinement and phrase-extraction heuristics (Koehn et al., 2003), the minimum error rate training (Och, 2003) for tuning, and 5-gram LMs with Kneser-Ney smoothing trained with SRILM (Stolcke, 2002).

[8] Both freely available from: `https://github.com/cocoxu/simplification/`

[9] We use the output of the original SBMT (Xu et al., 2016) and LightLS (Glavaš and Štajner, 2015) systems, obtained from the authors.

[10] Those cases in which the two annotators did not agree are additionally evaluated by a third annotator to obtain majority.

Approach	Training		LM		Changes		Scores		Rank
	Dataset	Size (sent)	Corpus	Size (sent)	Total	Correct	G	M	S
PBSMT	Wiki(Good+Partial)	284,499	Wiki	391,572	76	27 (35.5%)	4.09	3.31	0.26
PBSMT	Newsela(neighb. C3G-2s)+Wiki	593,947	Newsela+Wiki	766,446	81	38 (46.9%)	4.40	**3.84**	**0.30**
PBSMT	Newsela(all C3G-2s)+Wiki	764,572	Newsela+Wiki	766,446	87	42 (**48.3%**)	4.25	3.73	**0.30**
PBSMT	Newsela(neighb.-HMM)+Wiki	584,106	Newsela+Wiki	766,446	75	33 (44.0%)	4.20	3.65	0.20
s.o.t.a.	supervised SBMT (PPDB+SARI) (Xu et al., 2016)				**143**	49 (34.3%)	4.28	3.57	0.03
	unsupervised (LightLS) (Glavaš and Štajner, 2015)				132	35 (26.6%)	**4.47**	2.67	-0.01

Table 3: Extrinsic evaluation (PBSMT-based automatic text simplification systems vs. state of the art).

System	Output
Original	He advocates applying a user-centered design process in product development cycles and also works towards popularizing interaction design as a mainstream discipline.
PBSMT (Newsela neighb. C3G-2s + Wiki)	He **advocates a** user-centered design process in product development cycles and also works **for** popularizing interaction design as a mainstream discipline.
PBSMT (Newsela all C3G-2s + Wiki)	He **supports a** user-centered design process in product development cycles and also works **for** popularizing interaction design as a mainstream discipline.
PBSMT (Newsela HMM neighb. + Wiki)	He **advocates a** user-centered design process in product development cycles and also works towards popularizing interaction design as a mainstream discipline.
PBSMT (Wiki)	He advocates applying a user-centered design process in product development cycles and also works towards popularizing interaction design as a mainstream discipline.
SBMT (Xu et al., 2016)	He advocates **using** a user-centered design process in product development cycles and also works **for** popularizing *trade* design as a *whole field*.
LightLS	He *argues allowing* a user-centered design process in product development cycles and also works towards popularizing interaction design as a mainstream discipline.

Table 4: Outputs of different ATS systems (the correct changes/simplifications are shown in bold and the incorrect ones in italics).

(either *neighb. C3G-2l* or *all C3G-2l*) to the currently best sentence-aligned Wiki corpus (Hwang et al., 2015) in a standard PBSMT setup significantly[11] improves grammaticality (*G*) and meaning preservation (*M*), and increases the percentage of correct changes (Table 3). It also significantly outperforms the state-of-the-art ATS systems by simplicity rankings (*S*), meaning preservation (*M*), and number of correct changes (*Correct*), while achieving almost equally good grammaticality (*G*).

The level of simplification applied in the training dataset (*Newsela neighb. C3G-2s* vs. *Newsela all C3G-2s*) significantly influences G and M scores.

The use of the HMM-method for aligning Newsela (instead of ours) lead to significantly worse simplifications by all five criteria.

An example of the outputs of different ATS systems is presented in Table 4.

6 Conclusions

We provided several methods for paragraph- and sentence-alignment from parallel TS corpora, made the software publicly available, and showed that the use of the new sentence-aligned (freely available) Newsela dataset leads to state-of-the-art ATS systems even in a basic PBSMT setup. We also showed that lexically-based C3G method is superior to semantically-based methods (CWASA and WAVG) in aligning paraphraphs and sentences with 'heavy' simplifications (0–4 alignments), and that 2-step sentence alignment (aligning first paragraphs and then sentences within the paragraphs) lead to more correct alignments than the 'direct' sentence alignment.

[11] Wilcoxon's signed rank test, $p < 0.001$.

Acknowledgments

This work has been partially supported by the SFB 884 on the Political Economy of Reforms at the University of Mannheim (project C4), funded by the German Research Foundation (DFG), and also by the SomEMBED TIN2015-71147-C2-1-P MINECO research project.

References

Sandra Maria Aluísio and Caroline Gasperin. 2010. Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts. In *Proceedings of YIWCALA Workshop at NAACL-HLT 2010*. pages 46–53.

Stefan Bott and Horacio Saggion. 2011. An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*. ACL, pages 20–26.

Richard J. Evans. 2011. Comparing methods for the syntactic simplification of sentences in information extraction. *Literary and Linguistic Computing* 26(4):371–388.

Marc Franco-Salvador, Parth Gupta, Paolo Rosso, and Rafael E. Banchs. 2016. Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language. *Knowledge-Based Systems* 111:87 – 99.

Goran Glavaš and Sanja Štajner. 2015. Simplifying Lexical Simplification: Do We Need Simplified Corpora? In *Proceedings of the ACL&IJCNLP 2015 (Volume 2: Short Papers)*. pages 63–68.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning Sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of NAACL&HLT*. pages 211–217.

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 ACL*. pages 177–180.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the NAACL&HLT, Vol. 1*. pages 48–54.

Paul Mcnamee and James Mayfield. 2004. Character n-gram tokenization for European language text retrieval. *Information Retrieval* 7(1):73–97.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at International Conference on Learning Representations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*. pages 3111–3119.

Franz Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the ACL*. pages 160–167.

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

Gustavo Henrique Paetzold and Lucia Specia. 2016. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the 30th AAAI*.

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.

Gerard Salton and Michael J. McGill. 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.

Lucia Specia. 2010. Translating from complex to simplified sentences. In *Proceedings of the 9th PROPOR*. Springer Berlin Heidelberg, volume 6001 of *Lecture Notes in Computer Science*, pages 30–39.

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

David Vickrey and Daphne Koller. 2008. Sentence simplification for semantic role labeling. In *Proceedings of ACL&HLT*. volume 344–352.

Sanja Štajner, Hannah Bechara, and Horacio Saggion. 2015. A Deeper Exploration of the Standard PB-SMT Approach to Text Simplification and its Evaluation. In *Proceedings of ACL&IJCNLP (Volume 2: Short Papers)*. pages 823–828.

Sanja Štajner and Maja Popović. 2016. Can Text Simplification Help Machine Translation? *Baltic Journal of Modern Computing* 4(2):230–242.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics (TACL)* 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics* 4:401–415.