Cross-Sentence Transformations in Text Simplification

Fernando Alva-Manchego[1], **Carolina Scarton**[1] and **Lucia Specia**[1,2]
[1]Department of Computer Science, University of Sheffield
[2]Department of Computing, Imperial College London
{f.alva,c.scarton}@sheffield.ac.uk, l.specia@imperial.ac.uk

Abstract

Current approaches to Text Simplification focus on simplifying sentences individually. However, certain simplification transformations span beyond single sentences (e.g. joining and reordering sentences). In this paper, we motivate the need for modelling the simplification task at the document level, and assess the performance of sequence-to-sequence neural models in this setup. We analyse parallel original-simplified documents created by professional editors and show that there are frequent rewriting transformations that are not restricted to sentence boundaries. We also propose strategies to automatically evaluate the performance of a simplification model on these cross-sentence transformations. Our experiments show the inability of standard sequence-to-sequence neural models to learn these transformations, and suggest directions towards document-level simplification.

1 Introduction

Text Simplification (TS) aims to modify the content and structure of a text in order to make it easier to read and understand, while retaining its main idea. Current data-driven approaches for TS use sequence-to-sequence models to learn different simplification transformations altogether (Xu et al., 2016; Zhang and Lapata, 2017; Guo et al., 2018; Zhao et al., 2018), but are restricted to simplifying sentences one at a time, independently of wider context. Research on TS spanning multiple sentences (e.g. documents) is scarce and follows a similar approach: (i) to create candidate simplifications for each sentence in the text, and (ii) to use Integer Linear Programming to select which candidates to include in the output, satisfying global constraints based on document length (De Belder and Moens, 2010; Mandya et al., 2014) and information salience (Woodsend and Lapata, 2011). In this paper, we argue that document-level TS cannot be achieved solely by compression and content selection transformations over sentences that were simplified in isolation. We analyse professionally-produced document simplifications, and show that some transformations require information beyond sentence limits. Furthermore, we train standard sequence-to-sequence (seq2seq) neural models on simplification data, and evaluate their output to show their limitations when assessed at the document level.

2 Cross-Sentence Transformations

Our study is based on Newsela[1] (Xu et al., 2015), a parallel corpus of 1,130 news articles with up to five professionally-produced simplified versions each: the original text is version 0 and the most simplified version is 5. We randomly selected 4 articles and their 5 simplified versions (20 documents in total, since each article version is a document), and identified the transformations performed, focusing on those that depend on information beyond the sentence level.

Sentence Reordering. In Fig. 1, sentence 0-a was split into sentences 1-a and 1-c, and sentence 0-b was placed between the two. Current sentence simplification models would have placed the resulting splits in sequence, without any reordering.

[1]https://newsela.com/data, v.2016-01-29.

V0: (a) Facebook Chief Executive Mark Zuckerberg announced Tuesday that he plans to eventually donate 99 percent of the Facebook stock owned by him and his wife, Priscilla Chan, shares that are worth about $45 billion today. (b) That amount would make it one of the largest philanthropic commitments ever.
V1: (a) Facebook Chief Executive Mark Zuckerberg announced that he and his wife, Priscilla Chan, will donate 99 percent of their Facebook stock to charity. (b) Their promised gift would be one of the largest charitable donations ever made. (c) Together, the couple's shares are currently worth about $45 billion.

Figure 1: Example of sentence reordering.

Information Addition. In Fig. 2, sentences 1-b to 1-e have no equivalent in version 0. They were added to explain characteristics of communism (1-b and 1-c) and provide historical information from the trade embargo with Cuba (1-d and 1-e).

V0: (a) Gone, too, is the Hershey Social Club, [...] held dear by "Mister Hershey.". (b) "Everything has been destroyed," said Amparo DeJongh, 92, the first person born in the town and ... (c) "It's horrible what they have done," she said. (d) With U.S. businesses pushing harder than ever now against the Cuba trade embargo and angling ...
V1: (a) Gone, too, is the Hershey Social Club, [...] held dear by "Mister Hershey.". (b) Private business does not exist in communism. (c) Instead, the government controls business. (d) People from the United States who were running businesses in 1959 had to leave. (e) Then Washington put a trade embargo in place, which has prevented ... (f) With U.S. businesses pushing harder than ever now against the Cuba trade embargo and angling ...

Figure 2: Examples of information addition and content selection.

Sentence Joining. Editors join (parts of) sentences. In Fig. 3, for example, 1-b is split, and its first part is joined with 1-a to create 2-a.

V1: (a) At a later council meeting, some denounced Islam and Shariah, which is Islamic law. (b) One woman declared "Shariah law is Islam, and Islam's goal is to immigrate, assimilate and annihilate."
V2: (a) At a later council meeting, some denounced Islam and Shariah, including one woman who declared "Shariah law is Islam." (b) She said Islam's goal is to immigrate, become part of the wider society and then destroy it.

Figure 3: Example of sentence joining.

Content Selection. Editors sometimes delete (parts of) sentences. In Fig. 2, sentences 0-b and 0-c were removed in version 1. Also, in Fig. 4 sentence 0-a does not appear in version 1.

Anaphora Resolution. In Fig. 4, after removing 0-a, the personal pronoun *she* in 0-b was resolved to its antecedent entity *Elis de Cary Rojas*.

V0: (a) "We can't keep living like this" said Elis de Cary Rojas, who [...]. (b) She moved back to the town with her young daughter a few years ago ...
V1: (a) Elis de Cary Rojas moved back to the town with her young daughter a few years ago ...

Figure 4: Examples of content selection and anaphora resolution.

In order to quantify the manually identified cross-sentence transformations, we assume that to produce an article's simplified version, the editor uses its immediately preceding version, i.e., 0→1, 1→2, etc. Therefore, we extracted sentence alignments between adjacent articles' versions in the corpus using CATS (Štajner et al., 2018). For quantifying sentence reordering, we computed the number of sentences whose position in the document changed from its original version to its simplified version. For content selection and information addition, we calculated the number of unaligned original sentences and unaligned simplified sentences, respectively. For sentence joining, counting N-1 alignments suffices. Table 1 presents the counts for these first four transformations. For quantifying potential anaphora resolutions, we used Stanford CoreNLP (Manning et al., 2014) to extract the coreference chains in all documents for each version. Then, we counted how many of them contain coreferent pairs formed by entity mentions in different sentences (Table 2).

Table 1 shows that information addition is performed more frequently than the other three transformations. This could be because simplifying a text involves further explaining complex concepts. Although

Orig – Simp	ADD	DROP	REORD	JOIN
0 – 1	19,639	1,804	3,434	2,538
1 – 2	9,529	1,800	4,586	2,020
2 – 3	19,884	3,530	9,484	2,717
3 – 4	25,922	3,370	11,459	2,664
4 – 5	897	123	308	58
Total	75,871	10,627	29,271	9,997

Table 1: Cross-sentence transformations counts between adjacent original-simplified articles' versions.

Version	Coref. Chains	With Cross Mention Pairs
0	50,678	37,757 (74.5%)
1	46,493	35,960 (77.3%)
2	45,957	37,117 (80.8%)
3	42,645	35,984 (84.4%)
4	36,406	32,186 (88.4%)
5	652	601 (92.2%)

Table 2: Coreference chains statistics in the corpus.

less frequent, performing the other transformations impacts the structure and coherence of the document. As such, the large number of coreference chains (Table 2) signals that we need to pay attention when, for example, reordering or dropping a sentence, so as not to break the integrity of these coreferences.

3 Pseudo Document Simplification

Prediction Setting	BLEU↑	SARI↑
alignments	45.77	40.84
all	43.16	39.49

Table 3: Results of the sequence-to-sequence model.

We attempt to measure how a standard neural simplifier trained on sentence-level simplifications fairs at simplifying full documents. We split the Newsela corpus in train (80%), dev (10%) and test (10%) subsets, keeping all versions of each article in the same split. Since we are going to use a sentence-level model, we require aligned original-simplified sentences, which we obtained using CATS. As our simplification model, we used an encoder-decoder with attention as implemented in OpenNMT-py (Klein et al., 2017). For training, we followed Scarton and Specia (2018) by including "to-grade level" tags at the beginning of each sentence pair. Using these tags has shown to improve performance for neural sentence simplification models in the Newsela corpus. At test time, the model processes one document at a time and simplifies each of its sentences one by one. Then, these are placed sequentially to get the output for the document. We evaluated predictions in two settings: (1) *alignments*, where we only use sentences that have an aligned reference simplification; and (2) *all*, where we use all sentences in the original document, including those that are eventually dropped in its simplified reference. In order to measure "how close" these pseudo-simplified documents are from their references, we treat each of them as a "sentence", and use evaluation scripts from standard metrics: BLEU (Papineni et al., 2002) for grammaticality/meaning preservation, and SARI (Xu et al., 2016) for simplicity gain. Results are shown in Table 3. The scores in the *all* setting are lower than in the *alignments* one. Since the sentence-level model is neither dropping nor joining sentences, for example, it is harder to get closer to the reference simplification. However, these pseudo-document-level models could serve as baselines for the task.

4 Conclusion and Future Work

We have presented a study on cross-sentence transformations in professionally produced simplification corpora, and on the performance of standard neural seq2seq models on the task of simplifying full documents. Our analysis and experiments show that document simplification cannot be tackled by naively simplifying sentences in isolation, and then placing them in sequence. This is because extra-sentence decisions are needed. Moving forward, it is important to collect test sets and design evaluation metrics that are specific for each cross-sentence transformation, similar to Bawden et al. (2018) for evaluating coreference and coherence/cohesion in machine translation. In addition, it could be useful to study articles from other professionally produced simplification corpora, such as LiteracyWorks[2]. This could help to support our findings on cross-sentence transformations, and serve as a source for additional test data.

[2] http://literacynet.org/cnnsf/archives.html

References

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313. Association for Computational Linguistics.

Jan De Belder and Marie-Francine Moens. 2010. Text Simplification for Children. In *Proceedings of the SIGIR 2010 Workshop on Accessible Search Systems*, pages 19–26, Geneva. ACM.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Dynamic multi-level multi-task learning for sentence simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 462–476. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Angrosh Mandya, Tadashi Nomoto, and Advaith Siddharthan. 2014. Lexico-syntactic text simplification and compression with typed dependencies. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1996–2006, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, Pennsylvania. ACL.

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718. Association for Computational Linguistics.

Kristian Woodsend and Mirella Lapata. 2011. Wikisimple: Automatic simplification of wikipedia articles. In *Proceedings of the 25th National Conference on Artificial Intelligence*, pages 927–932, San Francisco, CA.

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*, 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.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Copenhagen, Denmark, September. Association for Computational Linguistics.

Sanqiang Zhao, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. 2018. Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3164–3173. Association for Computational Linguistics.

Sanja Štajner, Marc Franco-Salvador, Paolo Rosso, and Simone Paolo Ponzetto. 2018. Cats: A tool for customized alignment of text simplification corpora. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, may. European Language Resources Association (ELRA).