Not as Hard as it Looks: Exploring recent developments in Text Simplification—a paper reading report.

Reinelle Jan Bugnot
Nanyang Technological University
Singapore
G2304329L
bugn0001@e.ntu.edu.sg

ABSTRACT

Text simplification is a crucial technique with a clear objective: to make written documents more comprehensible and accessible, especially to individuals with diverse educational backgrounds and reading abilities, all while preserving the essence of the original content. This paper aims to explore three developments in the field text simplification by examining the innovative algorithms and model architectures the proponents implemented to perform the task and overcome existing challenges.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**.

KEYWORDS

NLP, text simplification, statistical machine translation

ACM Reference Format:
Reinelle Jan Bugnot. 2023. Not as Hard as it Looks: Exploring recent developments in Text Simplification—a paper reading report.. In *AI6122 (Text Data Management & Processing)*. S.Y. 2023-2024, Nanyang Technological University, Singapore, 4 pages.

1 INTRODUCTION

Simplified texts make it easier for a broader audience to engage with advanced content and is especially instrumental in assisting individuals with communication disorders such as aphasia, autism, or dyslexia, by presenting information in a more accessible and digestible format. Furthermore, text simplification plays a pivotal role as a pre-processing step in various downstream natural language processing tasks such as parsing, information extraction, and text summarization. By simplifying text before engaging in these subsequent tasks, the entire NLP process becomes more efficient and effective [1].

Text simplification is considered a sequence-to-sequence text generation problem in the field of natural language processing (NLP); similar to machine translation and paraphrasing. The core idea in all these tasks is to transform one piece of text into another while maintaining the essential meaning and content. There are two categories of text simplification based on the source text: sentence simplification and document simplification [1].

(1) **Sentence simplification** involves simplifying individual sentences within a text, typically one at a time. In this approach, the number of sentences in the input and output remains the same, but each sentence is made more understandable.
(2) **Document simplification**, on the other hand, focuses on reducing the number of sentences in the output text. This approach often involves summarizing or rephrasing paragraphs or sections to create a more concise and simplified version of the document.

While sentence simplification has received more attention in research and employs datasets designed for this purpose, document-level simplification offers distinct advantages, particularly in scenarios where a comprehensive understanding of multiple sentences at once is required. Aside from these, there are also two key aspects to consider in the domain of text simplification based on how the process is implemented: namely, lexical simplification and syntactic simplification [2].

(1) **Lexical simplification** involves the use of straightforward vocabulary and the inclusion of definitions that provide clear explanations in simple terms. It means replacing complex or less commonly understood words with their more accessible counterparts, ultimately reducing the cognitive burden on the reader.
(2) **Syntactic simplification**, on the other hand, centers on simplifying sentence structures and grammar. It often entails breaking down lengthy and convoluted sentences into shorter, more digestible ones. Additionally, it may involve rephrasing sentences to eliminate unnecessary complexity.

All these efforts collectively contribute to the overall goal of improving readability for any piece of text. In this report, I aim to explore three recent research works that cover these four areas and see how the proponents of each study introduce innovative ideas in each area to produce highly reliable and effective text simplification models.

2 NTS: NEURAL TEXT SIMPLIFICATION

The first attempt of using sequence-to-sequence neural networks to perform *lexical sentence simplification tasks* was published last 2017 by Nisioi, et. al. in their paper **Exploring Neural Text Simplification Models** [3]. Prior to their work, the task of text simplification was largely treated as a monolingual machine translation problem where a sentence or document is *translated* from its original form

Text Data Management & Processing, November 2023, Singapore

.

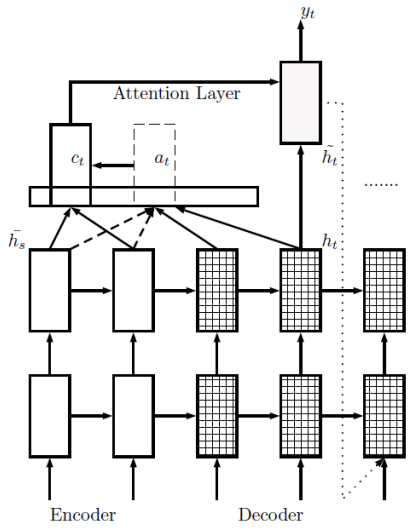

Figure 1: Architecture of the neural simplification model with global attention and input feeding [3]

into a simpler variant of the same language. Most of the state-of-the-art attempts at that time utilized classical statistical machine translation (SMT) systems such as phrase-based models [4], tree-based translation models [5], unsupervised word-embeddings [6], or re-ranking of the *n*-best outputs according to a dissimilarity metric [7].

In contrast to these SMT systems which often rely on external decoders, language models, and phrase tables, the authors emphasize that since neural networks are trained end-to-end, that they do not require these external components; rather, these features are extracted automatically. This shift towards end-to-end training simplifies and enhances the "translation" process, making it more efficient and flexible [3].

2.1 Implementation and Results

The authors of the research paper implemented the NTS model using the OpenNMT framework. Their architecture, shown in Fig. 1, included two Long Short-Term Memory (LSTM) layers in the encoder unit, and a decoder unit that employs global attention. On top of these is an attention layer from a context vector that uses the information learned within the hidden states of the LSTM layers to calculate alignment weights, which is an internal metric that the model uses to decide which of the previous learned hidden states will be passed onto the next step during input feeding. Alignment, in the context of NLP tasks, refers to word correspondence in a sequence-to-sequence model. The architecture implemented in this study, the authors argue, helps the model keep track of anterior alignment decisions. The proposed NTS model does not need to use character-based models to handle out-of-vocabulary (OOV) words; instead, the calculated alignment probabilities between the predictions and the original sentences can be used to retrieve the original words [3].

The NTS model was trained using a publicly available dataset based on English and Simple English Wikipedia (EW-SEW) and was

System	Output
NTS-w2v default	Perry Saturn (with terri) defeated Eddie Guerrero (with chyna) to win the WWF European Championship (8:10); Saturn pinned Guerrero after a diving elbow drop.
NTS-w2v SARI	Perry Saturn **pinned Guerrero to win the WWF European Championship.**
NTS-w2v BLEU	Perry Saturn pinned Guerrero after a diving **drop** drop.
NTS default	**He** (with terri) defeated Eddie Guerrero (with chyna) to win the WWF European Championship (8:10); Saturn pinned Guerrero after a diving elbow drop.
NTS BLEU/SARI	**He** defeated Eddie Guerrero (with Chyna) to win the WWF European Championship (8:10); Saturn pinned Guerrero after a diving elbow drop.

Figure 2: Neural Text Simplification Model Results [3]

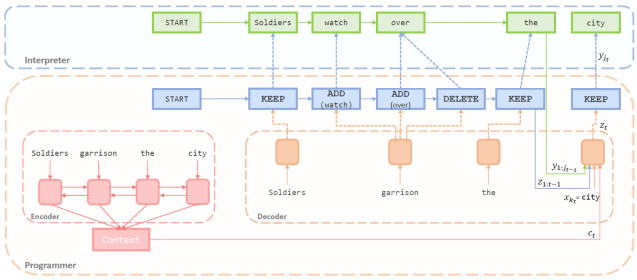

Figure 3: EditNTS Model Architecture [8]

shown to out perform state-of-the-art SMT techniques at the time in 3 different types of human-evaluated trials. An example output produced by the model is shown in Fig. 2, for the input sentence:

> *Perry Saturn (with terri) defeated Eddie Guerrero (with chyna) to win the WWF European Championship (8:10); Saturn pinned Guerrero after a diving elbow drop.*

3 EDITNTS: EXPLICIT SENTENCE EDITING

Machine translation (MT)-based models such as the ones discussed in section 2, demand that the model learns the simplification operations involved in transforming complex sentences into simpler ones, *implicitly*. While different MT-based strategies have produced promising results, these learned operations are mostly of very low frequency, so a large part of the original complex sentence often remains unchanged, as seen in Fig. 2.

The paper, **EditNTS: An Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing**, introduced by Dong, et. al. last 2019, instead proposes an alternative end-to-end *lexical sentence simplification* model that learns to *explicitly* generate the sequential edit operations required to produce a simple sentence from a complex sentence input, modelled after the way a human editor might perform sentence simplification [8].

The EditNTS model consists of two main components: the programmer and the interpreter. At each time step t, the programmer makes predictions regarding an edit operation z_t for the complex word x_{k_t}. This decision is influenced by several factors, including the words generated by the interpreter, the labels assigned by the programmer for the edit operation (which includes KEEP, DEL, and ADD operations), and a context attention vector derived from the entire complex sentence. Subsequently, the interpreter executes the edit operation z_t to produce the simplified token y_{j_t} and shares

WikiLarge	
Source	in 2005 , meissner became the second american woman to land the triple axel jump in national competition .
Output	meissner was the second american woman to land the triple axel jump .
Program	DEL DEL DEL KEEP ADD(was) DEL KEEP KEEP KEEP KEEP KEEP KEEP KEEP KEEP KEEP KEEP DEL DEL DEL KEEP
Reference	she is the second american woman and the sixth woman worldwide to do a triple axel jump .
WikiSmall	
Source	theodoros " thodoris " zagorakis -lrb- , born october 27 , 1971 in lyd -lrb- a village near the city of kavala -rrb- , is a retired greek footballer and was the captain of the greece national football team that won the 2004 uefa european football championship .
Output	zagorakis -lrb- born october 27 , 1971 is a former greek football player .
Program	DEL DEL DEL DEL KEEP KEEP DEL KEEP KEEP KEEP KEEP KEEP DEL DEL DEL DEL DEL DEL DEL DEL DEL DEL DEL DEL KEEP KEEP ADD(former) DEL KEEP ADD(football) ADD(player) DEL|DEL ... DEL KEEP
Reference	theodoros zagorakis -lrb- born 27 october , 1971 -rrb- is a former football player .
Newsela	
Source	schools and parent groups try to help reduce costs for low-income students who demonstrate a desire to play sports , she said .
Output	schools and parent groups try to help pay for low-income students .
Program	KEEP KEEP KEEP KEEP KEEP KEEP KEEP ADD(pay) DEL DEL KEEP KEEP KEEP DEL DEL DEL DEL DEL DEL DEL DEL DEL DEL KEEP
Reference	clark said that schools do sometimes lower fees for students who do n't have enough money .

Figure 4: EditNTS Sample Results [8]

the interpreter's context with the programmer to inform the next decision-making process. This collaborative interaction between the programmer and the interpreter is a key aspect of the EditNTS model [8]. The complete EditNTS architecture is shown in Fig. 3 and an example output is shown in Fig. 4.

4 SIMSUM: DOCUMENT SIMPLIFICATION

While the majority of current text simplification research has concentrated on sentence simplification, as indicated in the preceding sections, real-world applications often demand document-level simplification rather than sentence-level processing. This is primarily driven by the necessity to grasp the core ideas spanning several sentences simultaneously and rephrase them in a more straightforward vocabulary and grammar. Consequently, document-level text simplification holds promise for a broader range of applications compared to text simplification at the sentence level [1].

In this context, the paper titled **SimSum: Document-level Text Simplification via Simultaneous Summarization** authored by Blinova et al. shifts its focus to *document-level syntactic text simplification*. The paper introduces a two-stage model known as "SimSum," designed to address this specific aspect of text simplification [1]. It explores the challenges and strategies for simplifying the structure and content of documents to make them more accessible and understandable. This targeted approach aims to enhance the accessibility of complex textual content across various domains and applications.

"SimSum" is short for the two major components used in the architecture – a *Summarizer* transformer and a *Simplifier* transformer. The primary idea behind this architecture is that document-level simplification requires retaining the primary information from the original text (which is performed by the Summarizer), while restructuring the text to make comprehension easier (performed by the Simplifier). Fig. 5 shows the proposed framework that illustrates this idea [1].

The backbone model behind the implementation of SimSum is the T5 transformer model introduced by Raffel et. al. last 2019,

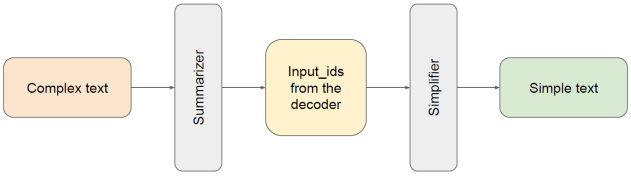

Figure 5: SimSum High-level Framework [1]

having shown high performances on a variety of NLP tasks which especially includes text summarization. A pre-trained version of the T5 model was used for both Summarizer and Simplifier model. The authors of the paper also created a version of SimSum that uses BART instead of T5, as reference. They eventually concluded that the T5 version of SimSum is better as it outperforms all benchmarked models for the document-level simplification tasks across majority of baselines.

5 CONCLUSION

The pursuit of improving text readability represents an ongoing and multifaceted NLP problem that has captured the attention of researchers across various domains. This report showcased three research works that introduced their own innovative ideas. Notably, these ideas build upon the foundations laid out by their predecessors, with a common objective: to enhance the comprehensibility of text. These contributions underscore the importance of considering various linguistic and cognitive factors when striving to make text more accessible to a wider audience, like non-native speakers, individuals with cognitive impairments, or those facing language barriers. This collective body of research reinforces the fundamental principle that clear and understandable communication is a vital aspect of bridging knowledge gaps and fostering inclusivity.

REFERENCES

[1] Sofia Blinova et al. "SIMSUM: Document-level Text Simplification via Simultaneous Summarization". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Ed. by Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki. Association for Computational Linguistics, 2023, pp. 9927–9944. DOI: 10.18653/V1/2023.ACL-LONG.552. URL: https://doi.org/10.18653/v1/2023.acl-long.552.

[2] Daniel Ferrés et al. "YATS: Yet Another Text Simplifier". In: *Natural Language Processing and Information Systems*. Ed. by Elisabeth Métais et al. LNCS 9612. Salford, UK: Springer, June 2016, pp. 335–342. DOI: 10.1007/978-3-319-41754-7_32.

[3] Sergiu Nisioi et al. "Exploring Neural Text Simplification Models". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Regina Barzilay and Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 85–91. DOI: 10.18653/v1/P17-2014. URL: https://aclanthology.org/P17-2014.

[4] Lucia Specia. "Translating from Complex to Simplified Sentences". In: *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language (PROPOR)*. Vol. 6001. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, pp. 30–39.

[5] Zhenghua Zhu, Daniel Berndard, and Iryna Gurevych. "A Monolingual Tree-based Translation Model for Sentence Simplification". In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. 2010, pp. 1353–1361.

[6] Goran Glavaš and Sanja Stajner. "Simplifying Lexical Simplification: Do We Need Simplified Corpora?" In: *Proceedings of the ACL & IJCNLP 2015 (Volume 2: Short Papers)*. 2015, pp. 63–68.

[7] Sander Wubben, Antal van den Bosch, and Emiel Krahmer. "Sentence Simplification by Monolingual Machine Translation". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL): Long Papers - Volume 1*. Association for Computational Linguistics, 2012, pp. 1015–1024.

[8] Yue Dong et al. "EditNTS: An Neural Programmer-Interpreter Model for Sentence Simplification through Explicit Editing". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3393–3402. DOI: 10.18653/v1/P19-1331. URL: https://aclanthology.org/P19-1331.