Journal of Theoretical and Applied Information Technology
30th November 2022. Vol.100. No 22
© 2022 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

THE LIMITS OF CONTEXT SIMPLIFICATION ON QA SYSTEM RESULTS

RACHID KARRA[1] , ABDELATIF LASFAR[2]
[1,2]LASTIMI Laboratory, Mohammadia School Of Engineers,
Mohammed V University in Rabat, Morocco

E-mail: [1]rachid.karra@est.um5.ac.ma, [2]abdelali.lasfar@est.um5.ac.ma

ABSTRACT

QA systems play a key role in human activities such as customer support, digital assistance in education, health, and public services. Our work is to use a QA system as a black box and see the effect of different simplification models on its results. The present study explores how far state-of-art simplification models can conserve text content. We measure text complexity with different linguistic metrics and meaning conservation with a BERT-based QA system score. Through evaluations, we measured text complexity and proved that context simplification as a multi-step simplification process gives better results in QA systems than ‘direct’ or ‘whole’ simplification. The proposed method has a better performance compared to automatic simplification. It is beneficial for a QA system with changeable contexts. As a task-oriented feature, choosing the convenient text simplification system should depend on its usefulness and the nature of the problem.

Keywords: *BERT, Linguistic analysis, QA system, Seq2Seq, Text simplification.*

1. INTRODUCTION

The purpose of simplifying text is to reduce its complexity and try to save as much information as the text conveys [1]. It is intended for a category of learners, such as K-12, foreign language students. It has undeniable benefits for some aphasic or dyslexic patients. The simplification process includes, among other things, summarizing, removing difficult content, reorganizing, and explaining. These techniques are used for several reasons such as:

- They are used by publishing professionals to simplify the text and thus meet school readability standards [2].
- To simplify texts for people with autistic disorders who sometimes have reduced reading abilities [3].
- Provide an accessible alternative for people who cannot read a language fluently due to text length and syntactic complexity [4].

Sentence simplification is also adopted in foreign language learning, where learning a new language should be done gradually without any linguistic sophistication. It helps non-native speakers and K-12 in text comprehension. Text simplification is used as a pre-processing step to facilitate text analysis and manipulation by parsers or other tasks such as information retrieval [5].

The tools used for simplification differ from the use of machine learning models or simple programmed models. Thus, we can explicitly model the simplification operators such as the insertion and deletion of words. [6] use splitting and deletion combined with sentence substitution and reordering [7]. We can also use NLP (Natural Language Processing) models [8] as sequence-to-sequence (seq2seq) models whose success largely depends on the quality and quantity of data used in training. In several simplification corpora, texts are classified on several levels according to their lexical and linguistic complexity. Newsela corpus contains sets of sentences: complex sentences and simplified text re-written by qualified editors ranked from level 0 to 4 [9]. Whereas Turk Corpus has eight ground-truth human simplifications for each complex sentence [10]. Selected Mturk workers were instructed not to split sentences and to conserve as much of the paragraph’s meaning and information as possible.

Automatic text simplification could help improve other NLP areas such as summarization, information extraction, and translation. In this study, we used two state-of-art Transformer models for sentence simplification. Transformer models use multi-layer, multi-head attention architecture. Compared to LSTM (Long Short-term Memory), the multi-head attention model would be able to process the entire input at once and choose the words to simplify the input sentence [11]. Despite its theoretical capability, LSTM can only memorize limited passage context in practice. Transformer-based tools for sentence simplification tasks perform better than LSTM-based models. The systems used such as T5, GPT-2, and BERT are based on Transformers.

2. RELATED WORKS

The early works concerning sentence simplification were rule-based, in particular aspects related to syntactic simplification [5]. For example, dividing complex sentences into several simpler ones [12]. He proposed a workflow based on five steps and several rules, each step has different tasks like POS tagging, structure or anaphoric preservation, or word disambiguation. Another difficult point in sentence simplification is text classification as complex or not, and therefore candidate to simplification or not. We use the Pointwise Mutual Information (PMI) statistical measure to decide whether a word is appropriate in a grade level context and thus whether the word is complex. On the other hand, Lexical simplification generally goes through Tokenization, selection of complex or less frequently used words using Zipf frequency or TF-IDF, and finally getting alternative words by ranking a list of synonyms from WordNet [13].

Recent years, text simplification was treated as a machine translation [14] and [7]. We have seen the increased use of machine learning techniques, especially for text simplification treated like other NLP tasks, such as machine translation or text summarization [15]. [16] proposes to solve the simplification problem with an encoder-decoder model coupled with a deep reinforcement learning framework. To solve the problem that seq-to-seq models have, it copies several words from the original sentences, [1] and [8] tried to develop a model to remedy this and give shorter and less complex results. [17] developed a model based on a multi-layer attention architecture and thus corrected one of the limitations of seq2seq models which favor frequent observations but neglect infrequently observed cases. [18] proposed an unsupervised and iterative Seq2Seq approach to address two limitations: The first one is that seq2seq models give little information about simplification operations and offer little control or adaptability to various aspects of simplification. The second is that they require a large amount of data. [19] introduces a hybrid approach that uses linguistically motivated rules with a Transformers-based paraphrase model.

While previous research efforts have developed new large language models on unlabeled datasets with billions of parameters, some studies have focused on datasets' data quality. Text simplification supporters argue that the syntactic and lexical changes used in simplification improve the text's readability and thus the reader's ability to understand and interact with a text [4]. Instead of developing LLMs trained on enormous amounts of text data, we should improve the quality of the training data and resources with which the language model interacts [15]. Many researchers are interested in the automatic improvement of data quality before feeding it into a model. The QA system mainly interacts with the context and question text. In e-learning for example, we can use small language models with the accuracy we need by adjusting context data quality.

To assess the level of simplification of the sentences, researchers can compare the text with a human reference of simplification or turn to reliable linguistic indicators such as L2 readability or syntactic complexity [20]. We enumerate categories of metrics like String Similarity (*BLEU, TER*), Flesch-based metrics (*FLE, FKGL*), and simplification (*SARI, SAMSA*) [21], [22]. BLEU index is less reliable for sentence simplification. [10] showed that the more simplified the sentence is like the original sentence, the higher the BLEU score. While concise sentences are one of the signs of a simplified text, BLUE shows little to non-existent correspondence with short sentences [22]. The authors combined some metrics to measure multiple aspects of phrases. [10] combines BLEU and FKGL to consider grammar and comprehension simplicity.

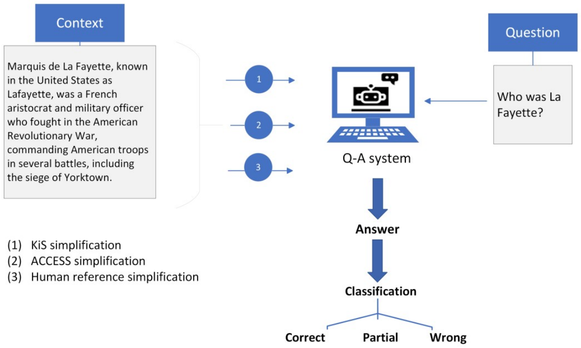
3. METHOD

In this section, we present experiments that compare our human reference simplification against the state-of-the-art simplification models alias ACCESS and KiS on a question-answering dataset. ACCESS (as a shorthand for AudienCe-CEntric Sentence Simplification) allows a parameterization mechanism that provides great control therefore users can condition simplifications on attributes such as length, lexical and syntactic complexity. KiS stands for Keep it Simple. It proposes a new approach to balance between well-formed sentences. The text should communicate the same information as the original one and be syntactically and lexically simpler.

QA system popularity is caused primarily by the wide adoption of recent technologies, the covid-19 pandemic, and the last decade's advances in artificial intelligence and neural networks [23]. Figure 1 shows an overview of our approach. First, we have a QA system based on BERT model; it answers a question from a given context. We adopted the same nomenclature and categorization of responses as in SQuAD [24].

We used a dataset of 81 questions, contexts, and the correct answers to evaluate the simplification of the contexts and the score of the QA system against the different simplifications. The paragraphs (contexts) are selected randomly from English Wikipedia. Each original paragraph is aligned with three simplifications: Human reference, KiS and ACCESS simplification. Then, we collect the three given answers and categorize them as correct, partial, or wrong [25]. We integrated context simplification and collection of responses in automated API testing using Postman. The classification of the obtained answers was done manually.

We have chosen two Transformer models trained on simplification tasks for the following reasons: they are the state-of-arts in simplification, the availability of source code, the reproducibility of results, and their excellent score in simplification tasks [26], [27].



Marquis de La Fayette, known in the United States as Lafayette, was a French aristocrat and military officer who fought in the American Revolutionary War, commanding American troops in several battles, including the siege of Yorktown.

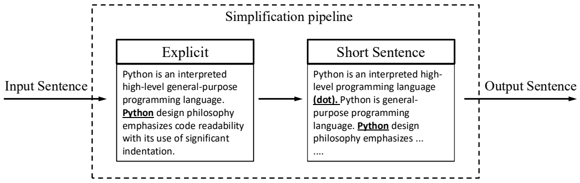
Who was La Fayette?

(1) KiS simplification
(2) ACCESS simplification
(3) Human reference simplification

Figure 1. The process of context simplification for QA system.

ACCESS supports four simplification parameters: Number of characters, Levenstein similarity, word lexical ranking, and depth dependency tree. We used the ACCESS model with NBchars 0.95, LevSim 0.75, and WordRank 0.75. [26] found the best simplification model score. They omitted the DepTreeDepth parameter probably because of an internal relationship between the parameters or the nature of the validation dataset. KiS is an unsupervised simplification model based on a rewarding method (k-SCST). It thus offers several simplified candidates for the same original text. Every GPT-2 simplification candidate receives a score according to simplicity, fluency, and salience. Salience gives the percentage of information in the original text covered by summarization or simplification [27]. In our simplification case, all the information of the original text must be restored in the corresponding output. The last parameter (binary flag) validates the simplifications or disqualifies them if they are not coherent, incorrect, or their compression rate is outside the interval [0.6, 1.5].

The ground-truth human simplification is treated as a two-stage process namely, explicit, and short sentences. The first step provides for every pronoun its explicit equivalent even if there is repetition or less text fluency. The second step uses original sentence splitting [25]. It is based mainly on sentence length. In practice, we did little rephrasing to preserve grammaticality. There was no drop of sentences or words, even if the authors noted redundancies. The goal of the reference simplification is to preserve the amount of information in the text and not its linguistic quality.



Python is an interpreted high-level general-purpose programming language. **Python** design philosophy emphasizes code readability with its use of significant indentation.

Python is an interpreted high-level programming language **(dot).** Python is general-purpose programming language. **Python** design philosophy emphasizes ... .

Figure 2. The architecture of the ground-truth human simplification system.

Secondly, we propose a simplification of context for a QA system to facilitate the extraction of answers. The study is inspired by text simplification widely used in language learning. The process of human reference simplification is shown in Figure 2.

GPT-2 is a transformer-based language model trained on a dataset of 8 million web pages and contains 1.5 billion parameters. GPT-2 aims to predict the next word with the previous words in the input sentence. GPT-2 proves that a model gets state-of-art scores in NLP tasks with unsupervised training [28]. The transformer is a sequence-to-sequence model composed of an encoder and a decoder (both of which are composed of several identical blocks). Each encoder block consists mostly of a multi-head self-attention module and a position-based feed-forward network (FFN).

Multiple metrics are proposed to measure text complexity in literature, especially in areas of linguistics and natural language processing. Indeed, having these measures available help text and natural language processing research [10]. These measures give us an idea of the text's semantic complexity, cohesion, and lexical simplicity. It allows the authors to have more understanding and information during the experimental research and investigate the correlation between different metrics. For readability, we lay on L2 Reading Index. It gauges the psycho-cognitive difficulty of reading. It is a more elaborate index, in the sense that it takes not only the descriptive aspects of the sentence but also words frequency, vocabulary complexity, semantics, and grammatical structure [20]. L2 Reading Index based on weighted values of CELEX frequency, Sentence Syntax Similarity, and Content Overlapping.

[22] criticizes the use of the BLEU index for simplification tasks. Indeed, BLEU is not adapted to structural simplifications and tends to well note the results with little modification.

SARI for « System output Against References and against the normal sentence » is an index used to measure text simplification. Whereas indices for translation tasks compares the input and output sentence of two different languages, SARI compares two sentences from one language. $SARI(T_{orig}, T_{ref}, T_{simpl})$ uses the original text, the reference simplification, and the simplified text to calculate the score. [10] demonstrated that SARI is the closest to human appreciation. The Flesch Reading Ease is one of the most adopted readability indices in the English language. It is a combination of the average length of the sentence (ASL) and the average number of syllables in a word (ASW).

$$FRE = 206.835 - (1.015 \text{ x } ASL) - (84.6 \text{ x } ASW) \quad (1)$$

FKGL metric (Flesch-Kincaid Grade Level) calculates the sentence length and word length, as a way to measure the complexity of a sentence.

$$FKGL = 0.39(NWSe) + 11.8\ NSyW - 15.59 \quad (2)$$

Where: NWSe = num words/num sentences

and NSyW = num syllables/ num words

The high value of FKGL indicates complex sentences. FKGL is inadequate for the human appreciation of a text or its simplification. On the other hand, SARI is less correlated to human judgment on facility and eloquence. The results of the QA system based on the BERT model using original context data have been analyzed.

4. RESULTS

In this section, we compare the impact of two automatic text simplification models on our QA system responses versus the human reference simplification. Also, we analyze some linguistic and simplification metrics.

Figure 3 shows the QA system scores for the Keep it Simple model. KiS model obtains a lower rate of correct results with 40.7% against 82.7% for manually changed contexts. On the other hand, the rate of incorrect answers after simplification with the KiS model is 55.6% against 9.9% for the human simplification method.

Figure 4 shows the QA system scores for the ACCESS model. The contexts changed with the ACCESS model also show a degradation of the rate of correct answers to 48.1% and those of incorrect answers to 45.7%. It should be noted that ACCESS records better results with 7.4 points more than KiS (48.1% against 40.7%) and 9.9 points more for wrong answers (45.7% against 55.6 %).

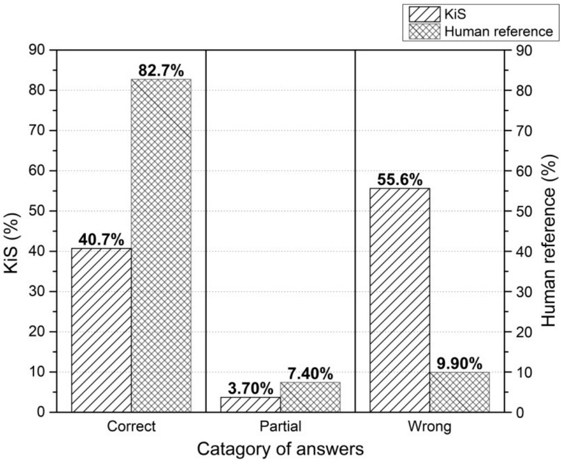


Figure 3. QA system responses for simplified contexts. Context simplified with KiS versus context simplified by workflow (Explicit and short sentence).

These results show that the manual simplification method obtains better results. This is in line with the observations raised by [27] who concluded that human simplification stills superior to those generated by models.

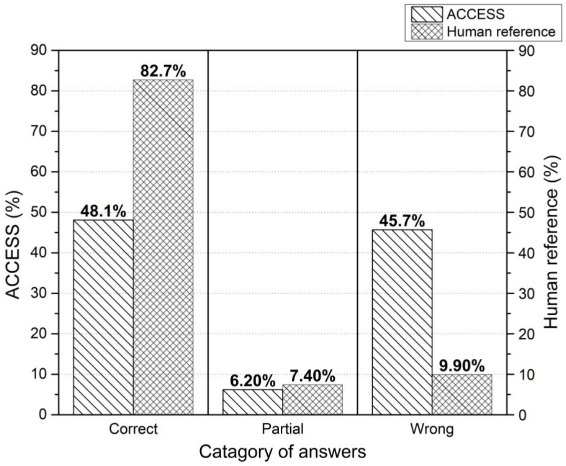


Figure 4. QA system responses for simplified contexts. Context simplified with ACCESS versus context simplified by workflow (Explicit and short sentence).

We analyzed the questions where the human simplification method recorded correct answers and KiS and ACCESS methods gave false answers with the COH-METRIX tool. Table 1 identifies some indicators with the three simplification experiments. Compared to automatic text simplification models, human simplification of [25] contains shorter sentences with more words, which has helped keep the contexts simple with the same amount of information.

The use of short sentences with fewer personal pronouns yielded more logical connectors which give consistency to the text. Finally, we extracted two of the indicators used to assess the text readability. KiS and ACCESS models record a score slightly lower than the manual simplification based on the two methods the “Explicit” and “short sentence.”

Table 1: Coh-Metrix metrics of the ground-truth human simplifications, KiS model and ACCESS model. Metrics are mean scores

Coh-Metrix metrics	Ground-truth	KiS	ACCESS
Sentence count, number of sentences	12.647	7.588	3.588
Word count, number of words	189.352	123.882	136.764
Sentence length, number of words, mean	15.662	16.833	42.559
Text Easability PC Narrativity, z score	-1.4707	-1.155	-0.875
Text Easability PC Syntactic simplicity, z score	0.827	0.722	-1.539
Text Easability PC Temporality, z score	0.252	0.177	-0.902
All connectives incidence	87.230	69.85	71.662
Left embeddedness, words before main verb, mean	0.520	0.602	0.621
Flesch Reading Ease	1.593	1.539	1.570

Sentence length is a descriptive metric that calculate the average number of word per sentence. More words means more complex syntax and difficulties to handle. Text Easability PC Syntactic simplicity, z score metric evaluates the two faces of words complexity and familiarity that make up the sentence (their number and their syntactic composition). All connectives incidence metric measure connectors’ impact. Connectors link sentence parts and T-units and make ideas organized and easy to understand. Finally, the Left embeddedness metric measures the number of words that precede the main verb of the sentence, and consequently the ability to memorize.

The context scores reported in the above-mentioned category are reinforced by the graphical comparison of the boxplots for the three methods. The first column concerns the manual simplification method, the second the KiS method, and the third the ACCESS method. Indeed, in most cases the different percentiles give the same results as the averages.

Table 2: We report the results for BLEU, SARI and FKGL metrics of the KiS and ACCESS models. ACCESS parameters are Nb chars (0.95), Lev-similarity (0.75) and Word ranking (0.75)

	BLEU ↑	SARI ↑	FKGL ↓
KiS	36.858	21.643	9.327
ACCESS	50.949	26.837	20.743

ACCESS scores best on the BLEU index at 50.949, higher than the KiS model. Although BLEU gives a result in agreement with the experiment, it remains less reliable for sentence simplification (see Table 2). The ACCESS model scores best on SARI 26.837. We observe KiS model can increase readability in terms of FKGL (9.327) compared to ACCESS paragraphs (20.743) shows a greater improvement in simple and short sentences than ACCESS [26].

The average number of words obtained by human reference simplification is 172, for KiS is 118, and 152 for ACCESS. We note that for the ACCESS model compression ratio is 75%, and the KiS model achieves a CR_{KiS} equal to 68%. Human reference simplification was the closest to the original text with a compression ratio of 102%. The manual simplification had the rules of not deleting anything and was framed by Explicit and short sentence methods. ACCESS is learning with a compression rate equal to 95%. Which is not always the right treatment for simplification.

KiS is trained for a length between 0.6 and 1.5. In our case, simplification must retain information, removing words or phrases necessarily remove pieces of information and result in a simplified context less ‘informative’ for the QA system. The KiS model scores best on FKGL (9.327) i.e., it generated shorter sentences than $ACCESS_{FKGL}$ (20.743). ACCESS shows high sentence length (39.153). KiS and Human-reference simplification perform better with 17.07 and 15.982, respectively. KiS_{FKGL} score was done at the expense of SARI (21.643). The authors observed that the ACCESS model outperformed the KiS model in terms of QA system correct answers despite having higher sentence length, but it achieves better performance in terms of BLEU and SARI. KiS lesser performance is partially due to its high compression rate, which necessarily reduces the amount of context information.

Table 3: Three examples for text simplification are provided. The bold part shows that the simplified context with KiS and ACCESS models deletes valuable information from the original context, whereas human simplification conserves it

Simplification model	Examples
KiS	… interfacing with the Amoeba, which is used by the show's main characters, and which is capable of handling any situation, from simple to complex, Van Rossum said.
ACCESS	…. with the Amoeba operating system as part of the company until his lead developer company 's began in **December 1989**, and after Van Rossum shoulder responsibility for the project, and Benthon started to work for the company.
Human reference	…. interfacing with the Amoeba operating system. Its implementation began in **December 1989**. Van Rossum …, as the lead developer, until 12 July 2018, when he announced his permanent vacation from his responsibilities as **Python's Benevolent Dictator For Life**, a title the Python community bestowed upon him to reflect his long-term ...

Table 3 shows the difference in simplification between the models. Information like the start date of the python implementation or the honorary title of Van Rossum has been removed from the simplification of KiS and ACCESS. QA system cannot achieve the same good score in KiS and ACCESS as the manual simplification. This negatively impacts the information quantity and consequently the amount of information that can be drawn from the context. The reference simplification is less invasive with a compression rate close to 1 and a good score in the QA system. It has a short sentence score of 15.982. Removing part of the context leads to a lack of information for the QA system, which influences its answers. Indeed, the QA system cannot get the span of text that answers a question. KiS and ACCESS have 'Deletions proportion' of 0.417 and 0.32, respectively. Compared to ACCESS, the KiS model prefers deletion-based simplification. ACCESS has more correct answers than KiS. The increase in deletion operations indicates that the models are less efficient.

5. DISCUSSION

It is beneficial for the QA system to use a conservative simplification pipeline for contexts that rarely delete or paraphrase sentences. Thus, the difficulty lies in categorizing content as important or irrelevant. In some cases, paraphrasing results in incorrect simplification and more syntactic complex outputs [17].

General text simplification driven by data supposes the availability of a huge amount of data. It is not the case for each "language dysfunctionality" like dyslexia or autism. The automatic text simplifications have made substantial progress at the syntactic, lexical, and grammatical levels. Today, they can paraphrase, add, and modify entire paragraphs. In recent years, they have become the state-of-the-art of simplification, thanks to their efficiency and satisfactory results. Nevertheless, the results of our study show their limits and confirm that they are not suitable for all cases. Thus, we need other solutions and alternatives like a pipeline simplification system.

Another limitation to using existing automatic tools for sentence simplification is that they tend to reduce the size of the paragraph by deleting words and reducing the information contained in the text. Sentence simplification has difficulty in learning less frequent rules; this may be due to datasets they trained on. Other studies suggest that lexical simplification by substitution is less invasive and augments the information entropy of the sentence, whereas lexical simplification by deletion reduces the words and therefore reduces the information entropy. We recommend restricting context simplification for QA systems to lexical simplification by substitution. Lexical simplification by substitution goes through preprocessing and tokenization. We should not resort to an automatic or multi-criteria simplification but provide for a simplification followed by the calculation of the amount of information that has been removed by this operation (information entropy) and set a limit (which may depend on the task and the degree of precision it requires).

Simplification cannot be done outside of its environment and its assigned objectives. It must be thought of as a task-based function. Indeed, each specific area needs a particular simplification. Aphasic individuals are sensitive to sentence construction, syntactic complexity, and static processing [29]. [30] relied on lexical replacement, morphological simplification and anaphora resolution for dyslexic children based on texts in French language. For second language learners, it is multi-meaning or less recurrent words that lead to confusion. Vocabulary simplification is suitable for them [31].

Text simplification in the case of the QA system must preserve the entropy of the context information. Therefore, repetitions and making the context clear are to be advocated rather than the deletion of sentences (less entropy). For these reasons, the authors believe that a scalable simplification pipeline, with different simplification types like explicit change, lexical, short sentence, or declarative sentences, is an effective alternative to automatic text simplifications. Again, having optimal deep learning simplification models that do only one type of simplification and are organized in workflow gives better performance for the QA system than automatic text simplification. Additional studies must be carried out to confirm the interest and effectiveness of modular simplification pipelines, according to domains and tasks, instead of automatic text simplification.

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel simplification method based on specialized simplification models grouped in workflows. It substantially outperformed automatic text simplification. We showed that even state-of-art simplification models can mislead the QA system because it reduces the quantity of paragraph information. Also, we have shown that context simplification (non-generalist), organized as an ensemble or pipeline, can benefit for a QA system. Despite the improvement of the simplification of the context on the performance of the QA system, the method has a non-exceedable limit. In this case, the used language model must therefore be larger, with more parameters and more training data.

The results of this study must be confirmed by using other language models, which are based on architectures different from that of BERT. In addition, the conservative simplification method should be compared to other QA-type datasets to confirm it. Future work involves developing a conservative QA system. We can adopt a two-step solution, the first is a pipeline for text simplification and the second is an estimator for the amount of conserved information. We believe that other types of semantic and lexical simplifications should be explored. Most simplifications are based on datasets specially designed for simplification and publicly available. It would be valuable to investigate the impact of simplification on diverse datasets. Needless to say, an appropriate QA system for each linguistic dysfunction, based on weighted metrics, is needed to help this specific population.

REFERENCES:

[1] R. Kriz *et al.*, “Complexity-Weighted Loss and Diverse Reranking for Sentence Simplification.” in *North American Association of Computational Linguistics,* Minneapolis, MN, USA, 2019, pp. 3137–3147.

[2] Y. Zhong, C. Jiang, W. Xu, and J. J. Li, “Discourse Level Factors for Sentence Deletion in Text Simplification,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 9709–9716.

[3] R. Evans, C. Orasan, and I. Dornescu, “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)*, Gothenburg, Sweden, 2014, pp. 131–140.

[4] W. M. Watanabe, A. C. Junior, V. R. Uzêda, R. P. de M. Fortes, T. A. S. Pardo, and S. M. Aluísio, “Facilita: reading assistance for low-literacy readers,” in *Proceedings of the 27th ACM international conference on Design of communication - SIGDOC ’09*, Bloomington, Indiana, USA, 2009, p. 29–36.

[5] R. J. Evans, “Comparing methods for the syntactic simplification of sentences in information extraction,” *Literary and linguistic computing*, vol. 26, no.4, 2011, pp. 371–388.

[6] Y. Dong, Z. Li, M. Rezagholizadeh, and J. C. K. Cheung, “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,* 2019, pp. 3393–3402.

[7] S. Narayan and C. Gardent, “Hybrid Simplification using Deep Semantics and Machine Translation,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, USA, vol. 1, 2014, pp. 435–445.

[8] Y. Zhao, L. Chen, Z. Chen, and K. Yu, “Semi-Supervised Text Simplification with Back-Translation and Asymmetric Denoising Autoencoders,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no.5, 2020, pp. 9668–9675.

[9] C. Jiang, M. Maddela, W. Lan, Y. Zhong, and W. Xu, “Neural CRF Model for Sentence Alignment in Text Simplification.” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 2020, pp. 7943-7960.

[10] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, “Optimizing Statistical Machine Translation for Text Simplification,” *Transactions of the Association for Computational Linguistics*, vol. 4, 2016, pp. 401–415.

[11] A. Vaswani *et al.*, “Attention is All you Need,” in *Advances in neural information processing systems 30,* 2017, pp. 5998–6008.

[12] A. Siddharthan, “An architecture for a text simplification system,” in *Language Engineering Conference, 2002. Proceedings*, Hyderabad, India, 2003, pp. 64–71.

[13] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, “Introduction to WordNet: An On-line Lexical Database,” *International journal of lexicography,* vol. 3, no.4, 1991, pp. 235–244.

[14] S. Wubben and E. Krahmer, “Sentence Simplification by Monolingual Machine Translation,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2012, pp. 1015–1024.

[15] F. Alva-Manchego, C. Scarton, L. Specia, " Data-Driven Sentence Simplification: Survey and Benchmark," Computational Linguistics, vol. 46, no.1, 2020, pp. 135–187.

[16] X. Zhang and M. Lapata, “Sentence Simplification with Deep Reinforcement Learning,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017, pp. 584–594.

[17] S. Zhao, R. Meng, D. He, A. Saptono, and B. Parmanto, “Integrating Transformer and Paraphrase Rules for Sentence Simplification,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018, pp. 3164–3173.

[18] D. Kumar, L. Mou, L. Golab, and O. Vechtomova, “Iterative Edit-Based Unsupervised Sentence Simplification.” in *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, 2020,pp. 7918–7928.

[19] M. Maddela, F. Alva-Manchego, and W. Xu, "Controllable Text Simplification with Explicit Paraphrasing." in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Mexico City, Mexico, 2021, pp. 3536–3553.

[20] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai, "Coh-Metrix: Analysis of text on cohesion and language," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no.2, pp. 193–202, 2004.

[21] J. P. Kincaid, J. R. P. Fishburne, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel." in *Naval Technical Training Command Millington TN Research Branch*, 1975, pp. 8–75.

[22] E. Sulem, O. Abend, and A. Rappoport, "BLEU is Not Suitable for the Evaluation of Text Simplification." in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* Brussels, Belgium, 2018, pp. 738–744.

[23] R. Karra and A. Lasfar, "Effect of Questions Misspelling on Chatbot Performance: A Statistical Study," in *International Conference on Digital Technologies and Applications, ICDTA 22,* Fez, Morocco, 2022, pp. 124–132.

[24] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proceeding of the Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2383–2392.

[25] R. Karra and A. Lasfar, "Impact of Data Quality on Question Answering System Performances," *Intelligent Automation & Soft Computing*, vol. 35, no.1, 2023, pp. 335–349.

[26] L. Martin, B. Sagot, É. de la Clergerie, and A. Bordes, "Controllable Sentence Simplification." in *Proceedings of the Twelfth Language Resources and Evaluation Conference,* Marseille, France, 2020, pp. 4689–4698.

[27] P. Laban, T. Schnabel, P. Bennett, and M. A. Hearst, "Keep it Simple: Unsupervised Simplification of Multi-Paragraph Text." in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 6365–6378.

[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," *OpenAI blog 1*, no.8, 2019, pp. 9.

[29] D. Pregla, P. Lissón, S. Vasishth, F. Burchert, and N. Stadie, "Variability in sentence comprehension in aphasia in German," *Brain and Language*, vol. 222, 2021, pp. 105008–105028.

[30] N. Gala, A. Tack, L. Javourey-Drevet, T. François, and J. C. Ziegler, "Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 1353–1361.

[31] M. Xia, E. Kochmar, and T. Briscoe, "Text Readability Assessment for Second Language Learners," in *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 2016, pp. 12–22.