TextSimplifier: A Modular, Extensible, and Context Sensitive Simplification Framework for Improved Natural Language Understanding

Sandaru Seneviratne[1], **Elena Daskalaki**[1], **Hanna Suominen**[1,2]
[1]The Australian National University (ANU) / Canberra, ACT, Australia
[2]University of Turku / Turku, Finland
{sandaru.seneviratne, eleni.daskalaki,
hanna.suominen}@anu.edu.au

Abstract

Natural language understanding is fundamental to knowledge acquisition in today's information society. However, natural language is often ambiguous with frequent occurrences of complex terms, acronyms, and abbreviations that require substitution and disambiguation, for example, by "translation" from complex to simpler text for better understanding. These tasks are usually difficult for people with limited reading skills, second language learners, and non-native speakers. Hence, the development of text simplification systems that are capable of simplifying complex text is of paramount importance. Thus, we conducted a user study to identify which components are essential in a text simplification system. Based on our findings, we proposed an improved text simplification framework, covering a broader range of aspects related to lexical simplification — from complexity identification to lexical substitution and disambiguation — while supplementing the simplified outputs with additional information for better understandability. Based on the improved framework, we developed TextSimplifier, a modularised, context-sensitive, end-to-end simplification framework, and engineered its web implementation. This system targets lexical simplification that identifies complex terms and acronyms followed by their simplification through substitution and disambiguation for better understanding of complex language.

1 Introduction

Limited reading or comprehension skills can hinder managing and maintaining a comfortable lifestyle in today's information society. Regardless of acquiring skills related to reading and comprehension over many years, sometimes understanding text can be challenging for, for example, people with limited reading skills, cognitive conditions like aphasia or dyslexia (Saggion et al., 2022), limited knowledge of technical domains, non-native speakers, and children (Kajiwara et al., 2013). Therefore, different methods have been introduced to assist with reading and comprehension of language, ranging from i) manual efforts of "translating" text to more understandable formats to ii) automated simplification methods (see Section 2).

Text simplification aims to modify the content and structure of complex text to output simpler text while preserving meaning. Commonly, the two main concepts associated with simplification are identified as readability and understandability. Even though these two concepts seem highly coupled, they address two different aspects of simplification (Shardlow, 2014): Readability focuses on how complex text can be converted to simple text to make it easier to read. In contrast, understandability is related to how much information a user can grasp from the text. Depending on the context and audience for which the text simplification is intended, the focus on improving the readability or understandability may differ.

Consequently, being sensitive to the different intents, researchers have introduced various methods for simplification (see Section 2): If the aim of the simplification is readability improvement, different methods focusing on the simplification of the syntactical structure have been proposed. These methods achieve simplification primarily by deleting, reordering, and splitting sentences to convert them to syntactically simpler formats so that the text is easier to read (Chandrasekar and Srinivas, 1997; Siddharthan, 2006). On the other hand, for understandability improvement, most methods focus on generating alternative substitutes for target complex words in text, focusing on the lexical simplicity of the text (Seneviratne et al., 2022c).

Improving the understandability of text benefits many audiences. For example, these understandability-focused simplification methods are helpful for non-native speakers and second-

Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability associated with RANLP-2023, pages 21–32,
held in Varna Bulgaria, Sept 7, 2023.
https://doi.org/10.26615/978-954-452-086-1_003

language learners to learn about new languages. Moreover, these methods can be helpful for students learning about new technical content or anyone who is not an expert in a specific technical domain. For example, domains like medical and scientific domains contain technical content, which is quite difficult for lay people to understand. Hence, extensive research has been done on the improved understandability of complex text (see Section 2).

Text simplification systems focusing on improved understandability of text explore different aspects related to simplification. For example, some methods investigate the complexities in text , (Pouran Ben Veyseh et al., 2021; Orlando et al., 2021), whereas others investigate the generation of alternatives for complex words (Azab et al., 2015; Paetzold and Specia, 2016) . Recent methods of text simplification rely on machine translation-based Sequence-to-Sequence (Seq2Seq) models for text simplification (Zhang and Lapata, 2017; Nisioi et al., 2017; Zhao et al., 2018; Maddela et al., 2021), which tackle both lexical and syntactic simplification of text. One of the limitations of Seq2Seq models is that they achieve simplification mainly by reducing the length of the sentences through the deletion of tokens which results in improved readability, however, at the cost of understandability (Maddela et al., 2021). Hence, when focusing on the understandability aspect of the text, modular approaches which tackle one subtask at a time may yield better outputs.

Generally, most practical simplification methods targetting lexical simplicity or understandability follow a modular approach with a pipeline proposed by Shardlow (2014). This pipeline comprises complex word identification, substitution generation, selection, and ranking methods for improved understandability. However, even though this pipeline has been adopted for many functional simplification systems, they only focus on complex words or phrases and simplification of them. For better understandability identifying other aspects that contribute to the complexity is essential. For example, in technical domains like medical or scientific, technical shorthand (i.e., acronyms or abbreviations for technical terms) are often used for ease of use and to avoid repetitions (Suominen et al., 2018). Hence, in such instances, shorthand identification and disambiguation of them is crucial for better understandability. Moreover, considering the complexities in text, in some instances, generating an alternative word or phrase may not be enough for accurate comprehension, thus requiring additional information.

The existing practical lexical simplification systems typically focus on one aspect of simplification, like addressing the complexity by acronyms (Pouran Ben Veyseh et al., 2021) or the ambiguity by the polysemic words (Orlando et al., 2021). In contrast, some systems rely on the pipeline by Shardlow (2014) and incorporate several components together (Bingel et al., 2018). Nevertheless, there are systems, which focus on both lexical and syntactic simplifications (Saggion et al., 2015; Ferrés et al., 2016). However, practical systems for lexical simplification at present have a limited coverage of components, thus requiring more comprehensive systems for practical lexical simplification.

In this paper, we present an improved text simplification framework targeting lexical simplification, extending the pipeline proposed by Shardlow (2014). It consists of the following four components: complex word identification, substitution generation, selection, and ranking. We report on a preliminary user study that we conducted to identify additional components required to enhance the simplification output for better understandability. Based on the user study, we have investigated and incorporated different components into the pipeline: i) an acronym identification module to address the complexities of shorthand, specifically acronyms, ii) an acronym disambiguation module to tackle the existence of multiple expansions for an acronym, and iii) an information module to supplement the outputs with more information for better understandability, together with iv) the conventional components. We have combined them as a pipeline to form both an improved framework and its implementation as a web-based system for lexical simplification, focusing on understandability. The proposed simplification system mainly addresses general-language and specialised (scientific/medical) text, due to the availability of resources and models.

2 Related Work

The earliest attempt to develop a text simplification system for practical use was made by Devlin and Unthank (2006), who introduced HAPPI — Helping Aphasic People Process Information, a web-based system to assist people with aphasia in reading online information. The system achieved

this by providing alternative words for complex words obtained through a database. The database consisted of psycholinguistic information about words like frequency and the familiarity of words used in the simplification process.

Text simplification systems for improved comprehension targeting lexical simplicity advanced in mid-2010s. Azab et al. (2015) introduced a text simplification system targeting second-language learners of the English language, with an interactive web interface for the users. The simplification was achieved by providing synonyms for complex words. Glavaš and Štajner (2015) proposed a resource-light, unsupervised lexical simplification system called LIGHT-LS. It relied on large regular text corpus for lexical simplification. A similar web interface to Azab et al. (2015) was introduced by Paetzold and Specia (2016). The tool was called Anita: An Intelligent Text Adaptation Tool and it relied on the LEXenstein framework by Paetzold and Specia (2015). Anita followed four steps in the simplification process where first candidate substitutes were produced based on a word embedding model followed by selection, ranking and replacement of the complex word. Additional information like synonyms and definitions were also provided in the system if a user requested it. Their method created user profiles intending to obtain users' feedback to improve the results. Bingel et al. (2018) introduced a text simplification tool called Lexi which also addressed obtaining users' feedback. The proposed system relied on the pipeline introduced by Shardlow (2014) and included complex word identification, substitution generation, selection, and ranking components. In addition, Lexi used users' feedback to personalise the experience to the target users.

Pioneering frameworks and systems for both lexical and syntactic simplification processes were also introduced over the years. Saggion et al. (2015) presented the Simplext project that effectively managed both lexical and syntactic simplification processes for Spanish. For lexical simplification, Simplext relied on a synonym-based and a rule-based simplification component, while for syntactic simplification handwritten computational grammars were used. Similarly, YATS by Ferrés et al. (2016) consisted of lexical and syntactic components to improve text readability and understandability for English. Its lexical simplification relied on a vector space model and word frequency simplicity measures to rank synonyms while its syntactic simplification used rule-based syntactic analysis and generation techniques based on part-of-speech tags and syntactic dependency information. Following a similar approach to YATS, a lexical simplification architecture for Spanish, Portuguese, Catalan, and Galician was introduced by Ferrés et al. (2017).

Focusing on the improved lexical simplicity of text, Orlando et al. (2021) introduced a word sense disambiguation system called AMuSE-WSD: An All-in-one Multilingual System for Easy Word Sense Disambiguation. The proposed system presented a web interface for word sense disambiguation in multiple languages. Addressing another aspect of lexical simplification, Pouran Ben Veyseh et al. (2021) proposed a web-based acronym identification and disambiguation system called MadDog. Its scope was entirely on the complexity added by acronyms.

Even though there have been several systems targeting lexical simplification in the recent past, most of these systems used the traditional lexical simplification pipeline by Shardlow (2014) for simplification, failing to consider the other essential components for improved understandability. Thus, it is important to explore beyond the traditional simplification steps when translating the research outputs into useful applications. Nevertheless, there has been extensive research in the domain of lexical simplification over the years that predominantly rely on transformer-based language models to improve the understandability of text (Saggion et al., 2022; Štajner et al., 2022).

3 User Study

The development of a functional text simplification system for practical use requires identifying what contributes to the complexity of the text, the aspects that should be considered, and the components that should be included. Thus, we conducted a user study to obtain user input on essential components for text simplification. Ethical approval (Protocol 2021/708) for the user study was obtained from the ANU Human Research Ethics Committee. The user responses were collected in a survey format.

3.1 Survey Process

To identify the essential components in a practical text simplification, we conducted a preliminary user study in the form of an online survey. We co-

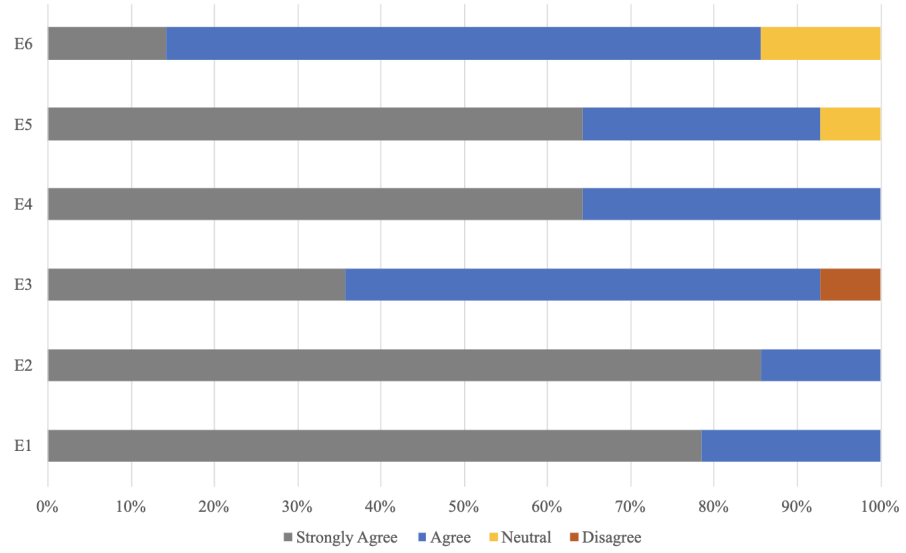

Figure 1: Evaluation results from the user study for all 14 participants. Labels of the $y - axis$ are as follows:
E1: Providing the correct expansion of shortened words is important for better understanding of unfamiliar acronyms.
E2: Inclusion of synonyms/similar substitutes for complex words is important for better understanding of complex text.
E3: Inclusion of additional information about words supplementing with definitions, links to more information can improve understandability of complex text.
E4: Systems that identify complex words and acronyms as well as provide substitutes, correct expansions, and additional information are useful.
E5: Grammatical structures and sentence structures can add complexity to text.
E6: Content simplification is more important than simplifying grammatical structures and sentence structures.

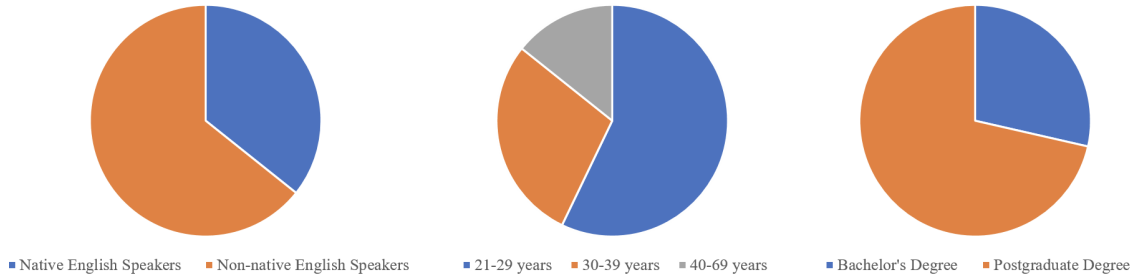

Figure 2: Participants' demographics based on their English-speaking background, age, and highest education.

created the survey questions included in this preliminary user study with user and health experience experts of the Our Health in Our Hands (OHIOH) health experience team (Figure 1). With the survey questions, we mainly targeted the complexities frequently found in complex medical text and the simplification of complex medical text. Each participant was asked to answer the set of survey questions, based on their experience, to identify what contributes to the complexity and components required for simplification. For each question, we provided four answer options (i.e., *strongly agree*, *agree*, *neutral*, and *disagree*).

3.2 Participants

We recruited participants from different English-speaking backgrounds, ages, and educational qualifications for the user study. In total, we recruited 14 participants, of which 9 were non-native English speakers and 5 were native English speakers. Most participants were in the age range of 21–29. In addition, the participants varied in their educational qualifications. For example, there were 4 participants with a bachelor's degree and 10 participants

with a postgraduate degree (Figure 2).

3.3 Evaluation Results

Seven out of the nine non-native participants expressed that they frequently or always encountered complex words in the text and found complex text challenging to understand. The native English speakers also indicated that they occasionally struggled to understand certain complex content, suggesting that despite their English-speaking background text can be complex. One reason might be that text from domains like medical or scientific domains we come across daily contains technical terms that are difficult for lay people to understand. Moreover, the exponential growth of information has resulted in a rapid increase of new words and terminologies that can be quite new to lay people.

All 14 participants, that is, both native and non-native English speakers, agreed that the inclusion of synonyms or alternative substitutes for complex words could improve text understandability. Moreover, through the survey, we asked the participants about acronyms and their associated complexities. We focused on the acronyms mainly because most technical domains often use shorthand for ease of use. This can result in complex text due to the availability of multiple possible expansions for one single acronym. All the participants agreed that identifying and disambiguating acronyms could improve text understandability.

Some of the previous systems provided supplementary information for complex text. Thus, in the survey, we asked the participants about their opinion on components to provide additional information. All the non-native participants agreed that including additional information could help the reader.

In addition to content simplification, we asked participants about the complexities of grammatical structures. The majority of the participants ($n = 13$) indicated that grammatical structures and sentence structures could contribute to the complexity of the text. Nevertheless, the results indicated that simplifying complex content is essential for understandability (Figure 1).

In the survey, we asked the participants their most commonly used methods to understand and simplify complex text. The majority of the participants indicated that they use internet searches, google, and dictionaries to find meanings of words. Some participants also indicated that they rely on Wikipedia for information needs relating to complex text. Regarding complex text in technical domains (e.g., medical), the participants stated that they would seek help from an expert in the field for clarification.

4 Proposed Framework

We proposed a modular text simplification framework for improved lexical simplicity based on feedback from the user study. The proposed framework extends beyond the conventional text simplification systems and pipelines and incorporates components targeting a much broader area of aspects related to lexical simplification.

Our work is founded on the pipeline by Shardlow (2014) with components for complex word identification, substitution generation, selection, and ranking. This can be converted into a pipeline with two components at a more abstract level forming it as a pipeline with complex word identification and lexical substitution, with the latter three components of the traditional pipeline falling under lexical substitution. The feedback from the user study indicated that acronyms also contribute to the complexity of text, and hence, we have incorporated an additional component for the acronym identification task. Following the acronym identification module, we have integrated a disambiguation module focusing on identifying the correct expansion of an acronym. Moreover, because the participants of the user study expressed the importance of a module to provide supplementary information for improved understandability, we have incorporated an information module into the pipeline.

Our proposed improved framework targets the understandability of natural language (Figure 3). It consists of 5 main modules: complex word identification, lexical substitution, acronym identification, acronym disambiguation, and information module.

5 System Design

We developed a modular, context-sensitive text simplification system based on the proposed framework focused on improved understandability (Figure 4) that we have made available at `http://130.56.247.69:8501/`. Each major component in the framework is a separate subfield in lexcal simplification. Hence, when translating the framework to the development stage, we have proposed new methods and relied on previous works for each component. The datasets used for experiments come

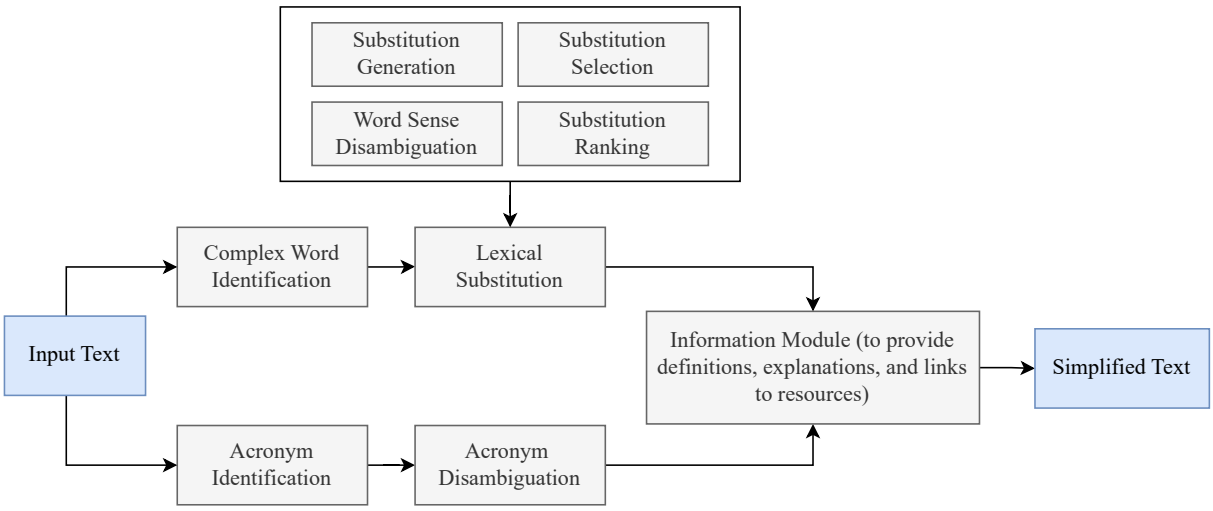

Figure 3: Our modular, extensible, and context-sensitive text simplification pipeline for improved understandability.

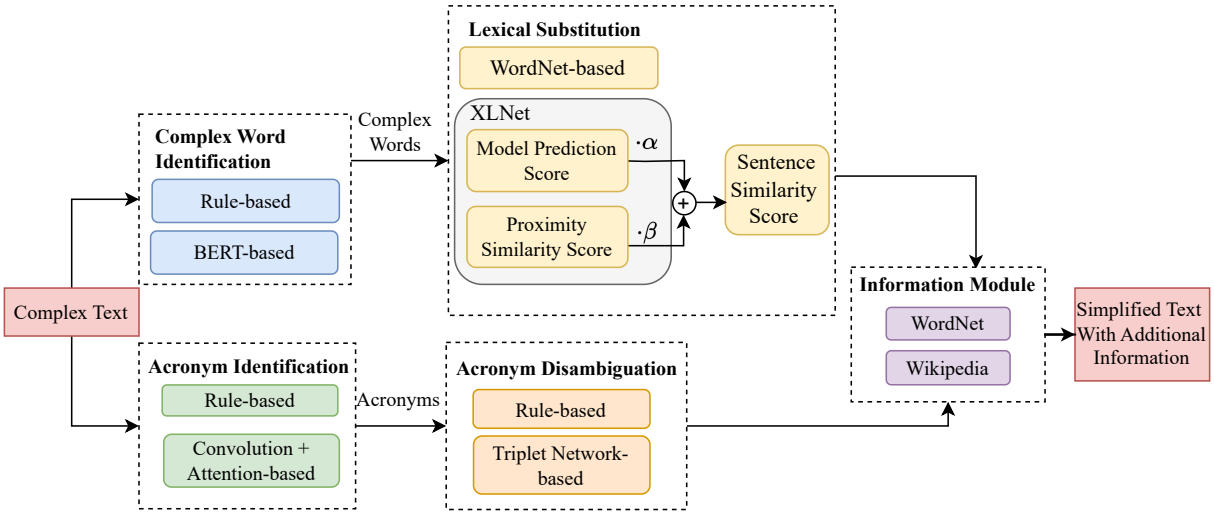

Figure 4: The system development of the proposed framework.

from general and specialised (scientific/medical) text. Its design and development consider dependencies of each of these components.

5.1 Complex Word Identification

In our system, we modeled identifying complex words as a token classification task, where the model predicts if the tokens in the input text are complex or not. We used the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model considering its effectiveness in many natural language processing tasks (Tenney et al., 2019; Qiang et al., 2020). The model was fine-tuned on the complex word identification dataset (Yimam et al., 2017) and achieved an `F1` score of 75%. In addition to the BERT-based model, a much simpler frequency-based complex word identification method, which used the frequency of a word per million words of English text based on Google Books Ngrams was included.

5.2 Lexical Substitution

The proposed toolkit has three lexical substitution methods. The first method generates WordNet-based synonyms for complex words while the other two methods are based on pre-trained language models proposed in Arefyev et al. (2020).

The lexical substitution method by Arefyev et al. (2020) relied on XLNet to produce layman-friendly alternatives for complex words by incorporating i) a model prediction score $P(w|c)$ where c is the context and w is any word from the XLNet vocabulary and ii) a proximity similarity score $P(w|x)$ where x is the target complex word as follows:

$$S_{\texttt{XLNet}} = \alpha P(w|c) + \beta P(w|x) \quad (1)$$

Method	P@1	
	LS07	CoInCo
BERT-based*	31.7	43.5
XLNet+embs	49.53	51.5
LexSubCon	51.7	50.5
CILex	53.38	55.73

Table 1: Results of substitution generation for LS07 and CoInCo datasets in %. We included reproduced results of the BERT-based substitution method (Zhou et al., 2019) by Michalopoulos et al. (2021) which is shown in *, reproduced the results of both i) XLNet+embs (Arefyev et al., 2020) and ii) LexSubCon (Michalopoulos et al., 2021). Our TextSimplifier uses the method in **bold**.

where α and β weigh the two scores.

Extending the XLNet-based method, we used CILex (Seneviratne et al., 2022a) a lexical substitution method that evaluates the added value of sentence context to ensure that the produced substitutions are semantically consistent and do not change the overall meaning of the sentences.

To evaluate the suitability of the possible candidates and their influence in the global context of the given sentence, we computed an additional score. Given a sentence s with a target word, we obtained an updated sentence (s') by replacing the target word with a possible substitution. For each possible substitution, a sentence similarity score was then calculated using cosine similarity using the sentence embeddings for the original sentence s and the updated sentence s':

$$S_{\texttt{sent}} = \texttt{cos}(s, s'). \quad (2)$$

The model score $S_{\texttt{XLNet}}$ and sentence similarity score $S_{\texttt{sent}}$ were linearly combined to rank and filter the final set of substitutions.

This proposed approach was tested on two publicly available datasets; Semeval 2007 task dataset (LS07) (McCarthy and Navigli, 2009) and the Concepts in Context (CoInCo) (Kremer et al., 2014) dataset. For both datasets, the proposed approach achieved state-of-the-art results in lexical substitution (Table 1).

5.3 Acronym Identification

We saw acronyms, formed from the first letters of words, as a sub-category of complex words in this study because of their contribution to the complexity. Hence, similar to complex word identification, we modeled acronym identification by defining the task as a token classification problem. To facilitate building the acronym identification model, we adopted the publicly available acronym identification dataset from the Scientific Document Understanding task, which consisted of labels for both acronyms and expansions (Pouran Ben Veyseh et al., 2020). For our experiments, we only considered the acronyms in the dataset. The model architecture consisted of convolutional neural networks and attention layers and achieved an `F1` score of 93.94% for the prediction of acronyms. Additionally, we also included a domain-independent rule-based acronym identification method proposed in Schwartz and Hearst (2002) in the toolkit which achieved an `F1` score of 92% .

5.4 Acronym Disambiguation

We modeled acronym disambiguation as a binary classification task to predict if the given expansion is the correct expansion or not for the corresponding acronym. We used a contrastive learning-based method to learn better representations of text and effectively disambiguate acronyms (Seneviratne et al., 2022b).

In the proposed approach, triplet loss

$$||f(x_i^a) - f(x_i^p)||_2^2 + \alpha < ||f(x_i^a) - f(x_i^n)||_2^2 \quad (3)$$

and triplet networks were leveraged to learn semantic differences among the different expansions of the same acronym through sentence triplet creation which included defining an anchor sentence, a positive sentence, and negative sentences. We defined the following process to sentence triplets: i) To obtain the list of anchor sentences, for each expansion of an acronym, we extracted a sentence randomly from the training subset of the data. The resultant set includes sentences with acronyms. ii) To obtain the positive sentences, we replaced the acronyms in the list of given sentences with their correct expansions. iii) To obtain negative sentences, first we obtained all the likely expansions of an acronym in the given sentence except for its correct expansion. Then, we obtained a list of sentences by replacing the given sentences' acronyms with these expansions. The resulting sentences were considered negative sentences. Consequently, we used the obtained anchor, positive, and negative sentences to train the triplet network-based architecture[1].

We validated the proposed approach both in the scientific and medical domains (Seneviratne

[1]More information on how the acronym disambiguation task was performed can be found at Seneviratne et al. (2022b).

Method	F1	
	SDU	MeDAL
Baseline method	59.73	44.39
Span prediction method	84.24	74.91
Triplet Network-based	85.70	75.19

Table 2: Results of acronym disambiguation for the validation data of SDU dataset and test data of MeDAL datasets in %. We included results reproduced using the i) frequency-based baseline method by Veyseh et al. (2020), ii) span prediction method by Singh and Kumar (2021), and iii) triplet network-based method by Seneviratne et al. (2022b). Our TextSimplifier is uses the method in **bold**.

et al., 2022b) using two publicly available datasets; acronym disambiguation dataset from Scientific Document Understanding Task (SDU) (Pouran Ben Veyseh et al., 2020) and a part of Medical Abbreviation Disambiguation Dataset (MeDAL) (Wen et al., 2020). Triplet Network-based method gave comparable performance as the baseline for both the datasets (Table 2). Furthermore, we included the domain-independent frequency-based baseline method by Pouran Ben Veyseh et al. (2020) in TextSimplifier toolkit.

5.5 Information Module

We engineered our Information Module to collect additional information related to predicted complex words and acronym expansions. Each complex word and acronym expansion was linked to its corresponding web page from Wikipedia. Web pages from both English and simple Wikipedia were used for this purpose. We envisioned users clicking on links to obtain further information. For better text understanding, definitions obtained from WordNet and disambiguated using sentence-Transformers (Reimers and Gurevych, 2019) were provided and integrated as a component in the system.

6 Discussion

In this paper, we have proposed a text simplification framework targeting improved lexical simplicity/language understandability using the feedback obtained through a user study on text complexities. Based on the feedback, we have extended the conventional lexical simplification pipeline to incorporate additional components essential for natural language understanding. As a result, we have derived a framework of complex word identification, lexical substitution, acronym identification, and acronym disambiguation components followed by an information module to supplement the simplified output.

Even though the typical lexical simplification systems focus only on the complexities of complex words and phrases, the evaluation results of the user study indicated that the acronyms contribute to the complexity of the text. One reason might be that acronyms are heavily used in technical domains like scientific and medical domains we come across daily. Moreover, the exponential growth in information has increased the use of acronyms. Hence, it is essential to identify and disambiguate them to determine the correct expansion corresponding to the meaning of the context and incorporate the relevant components in simplification pipelines. The results from the user study also indicated the importance of providing additional information related to the complexities in the text to improve understandability, thereby helping the readers grasp the knowledge effectively. Therefore, it is crucial to incorporate components that supplement the simplified versions of complex text.

In our proposed text simplification framework, we have integrated multiple components that all relate to lexical simplification. We have validated and assessed each component separately to ensure their effectiveness. Nevertheless, because each task was trained using datasets from different sources, this could potentially impact the final output. Therefore, exploring the compatibility of these separate models within a unified system is crucial. Moreover, the end-to-end pipeline as a whole was not evaluated. Thus, as future work, we expect to create datasets that provide annotations for each important task in a consistent manner, which could further enhance the effectiveness of text simplification methods. Given these challenges, the output generated by the complete pipeline has not been evaluated using a simplicity metric in this study.

The proposed simplification framework incorporates additional components required for improved language understandability compared to existing simplification systems. It also follows a modular or task-based approach in tackling different aspects related to simplification, which is much more explainable compared to models that rely on one black-box architecture for the simplification task. Moreover, its modular architecture eases the integration of new modules addressing other aspects of simplification and new components for each module in the

Input	The purpose of RL is for the agent to learn an optimal, or nearly-optimal, policy that maximizes the reward function .
TextSimplifier	The purpose of **RL (reinforcement learning)** is for the agent to learn an optimal, or nearly-optimal, policy that maximizes the **reward (payoff, incentive, benefit)** function (reinforcement learning: https://en.wikipedia.org/wiki/Reinforcement_learning reward: https://simple.wikipedia.org/wiki/Reward, reward: act or give recompense in recognition of someone's behavior or actions)
MadDog	The purpose of **RL (Reward Learning)** is for the agent to learn an optimal , or nearly - optimal , policy that maximizes the reward function .
Lexi (Hero)	The purpose of RL is to learn the best policy. The best policy will give the best reward.

Table 3: Comparison with existing toolkits; Lexi (Bingel et al., 2018), MadDog (Pouran Ben Veyseh et al., 2021).

framework. These features of the framework facilitate uncomplicated translation of the framework to the functional systems.

This paper has proposed a text simplification framework targeting the improved understandability of complex text. However, the evaluation results from the user study indicated the complexities of grammatical and sentence structures; hence, incorporating components for syntactic simplification is important. Therefore, future work is welcome to explore the addition of syntactical simplification components along with other modules that can be incorporated into the current framework for improved understandability.

Limitations

The main focus of the proposed user study is limited to the the simplification of complex words and acronyms. This could further be extended to incorporate the role of coherence/cohesion or the impact of syntactic complexity on understanding. Moreover, the participants of the user study are all well-educated even though some have English as their second language. Thus, the feedback could not be representative of the general audience requiring simplification of complex words.

We used the proposed framework for the development of a sample prototype system as a first step towards translating research into the real world. However, developing a text simplification system for practical use requires consideration of many different aspects, thus, is more complex. For example, given that the system aims to assist readers in improving their understandability, the system should have accurate and fast responses. This requires further validation of the outputs from the models to ensure that they do not generate incorrect responses, misinforming the readers. Moreover, the current methods rely heavily on deep learning models; hence, the efficient integration of the models is required. Our current prototype system is in early stages of development and hence it is advisable to be aware of the risks.

Ethics Statement

Ethical approval (Protocol 2021/708) was obtained from the ANU Human Research Ethics Committee for the user study. According to the National Statement on Ethical Conduct in Human Research (2007) — Updated 2018 (National Health and Medical Research Council, 2018), a new ethics approval was not required, and, to the best of our knowledge, all the datasets used were created ethically.

Acknowledgement

This research was funded by and has been delivered in partnership with Our Health in Our Hands (OHIOH), a strategic initiative of the ANU, which aims to transform health care by developing new personalised health technologies and solutions in collaboration with patients, clinicians and healthcare providers. We gratefully acknowledge the funding from the ANU School of Computing and OHIOH for the first author's PhD studies and the related travel. We also thank Dr. Nicola Brew-Sam of OHIOH Health Experience Team for her valuable insights on the user study. This work was supported by computational resources provided by the Australian Government through the National Computational Infrastructure (NCI) under the ANU Merit Allocation Scheme. We wish to thank NCI Australia for providing cloud resources for the project *ny83*, to host the demonstration system.

Lay Summary

Understanding language can be difficult due to complex words, acronyms, and abbreviations. People with limited reading skills, non-native speakers, and those learning a new language find it challenging. To simplify text, at present, automated text simplification methods are used. In this paper, we introduced a text simplification system that uses natural language processing and machine learning techniques. We conducted a user study to figure out different components important in text simplification systems. The proposed text simplification system first identifies complex terms that might confuse readers and then replaces them with simpler words. This TextSimplifier system also identifies acronyms or shortened words in text, provides their long expansion, and gives more information for complex words and acronyms to make things even clearer. This helps make information open to everyone, no matter their language skills.

References

Nikolay Arefyev, Boris Sheludko, Alexander Podolskiy, and Alexander Panchenko. 2020. Always keep your target in mind: Studying semantics and improving performance of neural lexical substitution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1242–1255, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mahmoud Azab, Chris Hokamp, and Rada Mihalcea. 2015. Using word semantics to assist English as a second language learners. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 116–120, Denver, Colorado. Association for Computational Linguistics.

Joachim Bingel, Gustavo Paetzold, and Anders Søgaard. 2018. Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Raman Chandrasekar and Bangalore Srinivas. 1997. Automatic induction of rules for text simplification. *Knowledge-Based Systems*, 10(3):183–190.

Siobhan Devlin and Gary Unthank. 2006. Helping aphasic people process online information. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 225–226.

Daniel Ferrés, Montserrat Marimon, Horacio Saggion, and Ahmed AbuRa'ed. 2016. Yats: yet another text simplifier. In *Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings 21*, pages 335–342. Springer.

Daniel Ferrés, Horacio Saggion, and Xavier Gómez Guinovart. 2017. An adaptable lexical simplification architecture for major Ibero-Romance languages. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 40–47, Copenhagen, Denmark. Association for Computational Linguistics.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, Beijing, China. Association for Computational Linguistics.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. 2013. Selecting proper lexical paraphrase for children. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 59–73, Kaohsiung, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What substitutes tell us - analysis of an "all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549, Gothenburg, Sweden. Association for Computational Linguistics.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. 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*, pages 3536–3553, Online. Association for Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2009. The english lexical substitution task. *Language resources and evaluation*, 43(2):139–159.

George Michalopoulos, Ian McKillop, Alexander Wong, and Helen Chen. 2021. Lexsubcon: Integrating knowledge from lexical resources into contextual embeddings for lexical substitution. *arXiv preprint arXiv:2107.05132*.

National Health and Medical Research Council. 2018. National Statement on Ethical Conduct in Human Research (2007). https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018. [Online; accessed 06-January-2022].

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Riccardo Orlando, Simone Conia, Fabrizio Brignone, Francesco Cecconi, and Roberto Navigli. 2021. AMuSE-WSD: An all-in-one multilingual system for easy Word Sense Disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 298–307, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2015. LEXenstein: A framework for lexical simplification. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 85–90, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Gustavo Paetzold and Lucia Specia. 2016. Anita: An intelligent text adaptation tool. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 79–83, Osaka, Japan. The COLING 2016 Organizing Committee.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Walter Chang, and Thien Huu Nguyen. 2021. MadDog: A web-based system for acronym identification and disambiguation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 160–167, Online. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Quan Hung Tran, and Thien Huu Nguyen. 2020. What does this acronym mean? introducing a new dataset for acronym identification and disambiguation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3285–3301, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jipeng Qiang, Yun Li, Zhu Yi, Yunhao Yuan, and Xindong Wu. 2020. Lexical simplification with pretrained encoders. *Thirty-Fourth AAAI Conference on Artificial Intelligence*, page 8649–8656.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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 (TACCESS)*, 6(4):1–36.

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2022. Findings of the TSAR-2022 shared task on multilingual lexical simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 271–283, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Ariel S Schwartz and Marti A Hearst. 2002. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Biocomputing 2003*, pages 451–462. World Scientific.

Sandaru Seneviratne, Elena Daskalaki, Artem Lenskiy, and Hanna Suominen. 2022a. CILex: An investigation of context information for lexical substitution methods. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4124–4135, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Sandaru Seneviratne, Elena Daskalaki, Artem Lenskiy, and Hanna Suominen. 2022b. m-networks: Adapting the triplet networks for acronym disambiguation. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 21–29, Seattle, WA. Association for Computational Linguistics.

Sandaru Seneviratne, Elena Daskalaki, and Hanna Suominen. 2022c. CILS at TSAR-2022 shared task: Investigating the applicability of lexical substitution methods for lexical simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 207–212, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4:77–109.

Aadarsh Singh and Priyanshu Kumar. 2021. Scidr at sdu-2020: Ideas-identifying and disambiguating everyday acronyms for scientific domain. In *In SDU@AAAI-21*.

Sanja Štajner, Daniel Ferrés, Matthew Shardlow, Kai North, Marcos Zampieri, and Horacio Saggion. 2022. Lexical simplification benchmarks for english, portuguese, and spanish. *Frontiers in Artificial Intelligence*, 5:991242.

Hanna Suominen, Liadh Kelly, Lorraine Goeuriot, et al. 2018. Scholarly influence of the conference and labs of the evaluation forum ehealth initiative: review and bibliometric study of the 2012 to 2017 outcomes. *JMIR research protocols*, 7(7):e10961.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Thien Huu Nguyen, Walter Chang, and Leo Anthony Celi. 2020. Acronym identification and disambiguation shared tasks for scientific document understanding. *arXiv preprint arXiv:2012.11760*.

Zhi Wen, Xing Han Lu, and Siva Reddy. 2020. MeDAL: Medical abbreviation disambiguation dataset for natural language understanding pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 130–135, Online. Association for Computational Linguistics.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017. CWIG3G2 - complex word identification task across three text genres and two user groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 401–407, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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 584–594, Copenhagen, Denmark. 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, Brussels, Belgium. Association for Computational Linguistics.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.