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

Australian National University

Sandaru Seneviratne, Eleni Daskalaki, Hanna Suominen

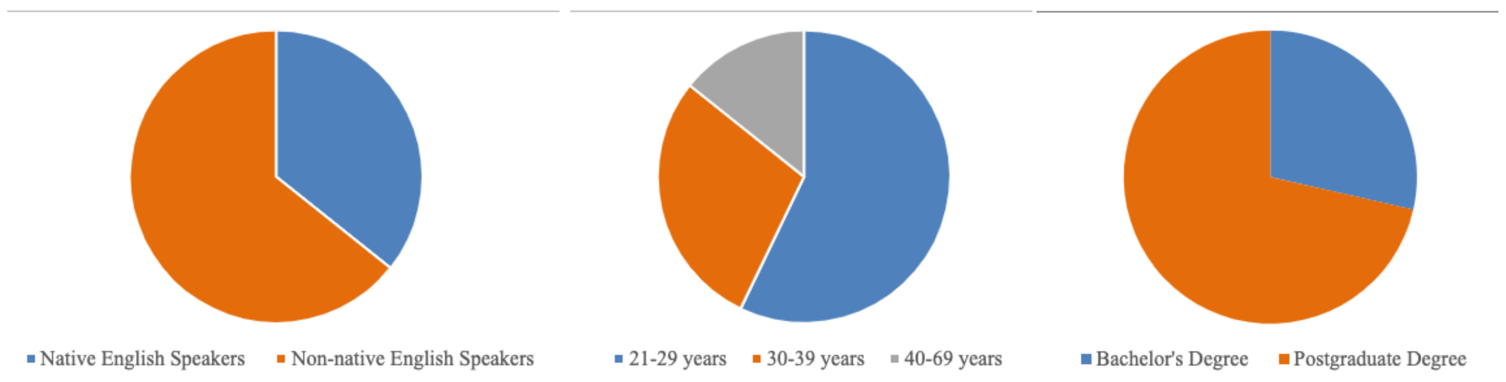
sandaru.seneviratne@anu.edu.au

Abstract

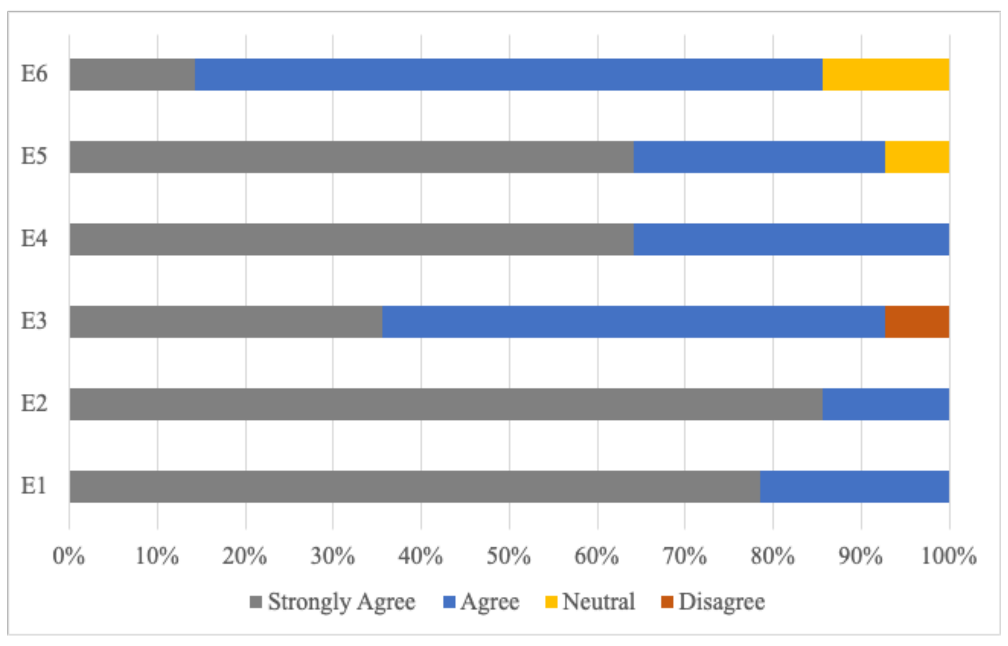
- Natural language is often ambiguous with frequent occurrences of complex terms, acronyms, and abbreviations that require simplification.
- When using automated text simplification methods, it is important to identify essential components in a text simplification system. Thus, we conducted a user study.
- Based on the user study, we propose a text simplification framework targeting lexical simplification.
- The system extends the text simplification pipeline proposed by (Shardlow, 2014) [1].
- Our framework called TextSimplifier consists of the following components.
  1. complex word identification
  2. lexical substitution
  3. acronym identification (new)
  4. acronym disambiguation (new)
  5. information module (new)

User Study

- After obtaining the proper ethics approvals and research permissions, we recruited participants of different English-speaking backgrounds, ages, and educational qualifications for the user study.
- We conducted an online survey to obtain user input on essential components and aspects in a text simplification system.
- We co-created the survey questions of this preliminary user study with user and health experience experts, mainly targeting the complexities frequently found in complex medical text.



Participants' demographic information: English-speaking background, age, and education.



Evaluation results from the user study for all the 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 compared to simplifying grammatical structures and sentence structures.

Method

- TextSimplifier is modular and has 5 disjoint modules.
- Each major module is a separate subfield in text simplification and developed separately.
- Its design and development consider dependencies to ease extensions.
- Preliminary system demo: http://130.56.247.69:8501/

Complexity Identification

Complex Word Identification

- Complex Word Identification Task dataset 2018 – News, WikiNews, Wikipedia articles.
- BERT-based method (F1 score: 75%)
- Frequency of a word per million words of English text based on Google Books Ngrams

Acronym Identification

- Acronym identification dataset from Scientific Document Understanding Task – scientific papers.
- CNNs+attention (F1 score: 93.94%)
- Rule-based method [2] (F1 score: 92%)

Lexical Substitution

- XLNet model was used to compute a model prediction score and an embedding similarity score. We defined S_{XLNet} as follows:

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

- **Sentence similarity score**: to ensure that the candidates fit in the global context of the sentence [3].

$$S_{sent} = \cos(s, s')$$

$$S = \gamma S_{XLNet} + \delta S_{sent}$$

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**

Acronym Disambiguation

- Triplet-network-based method [4].

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

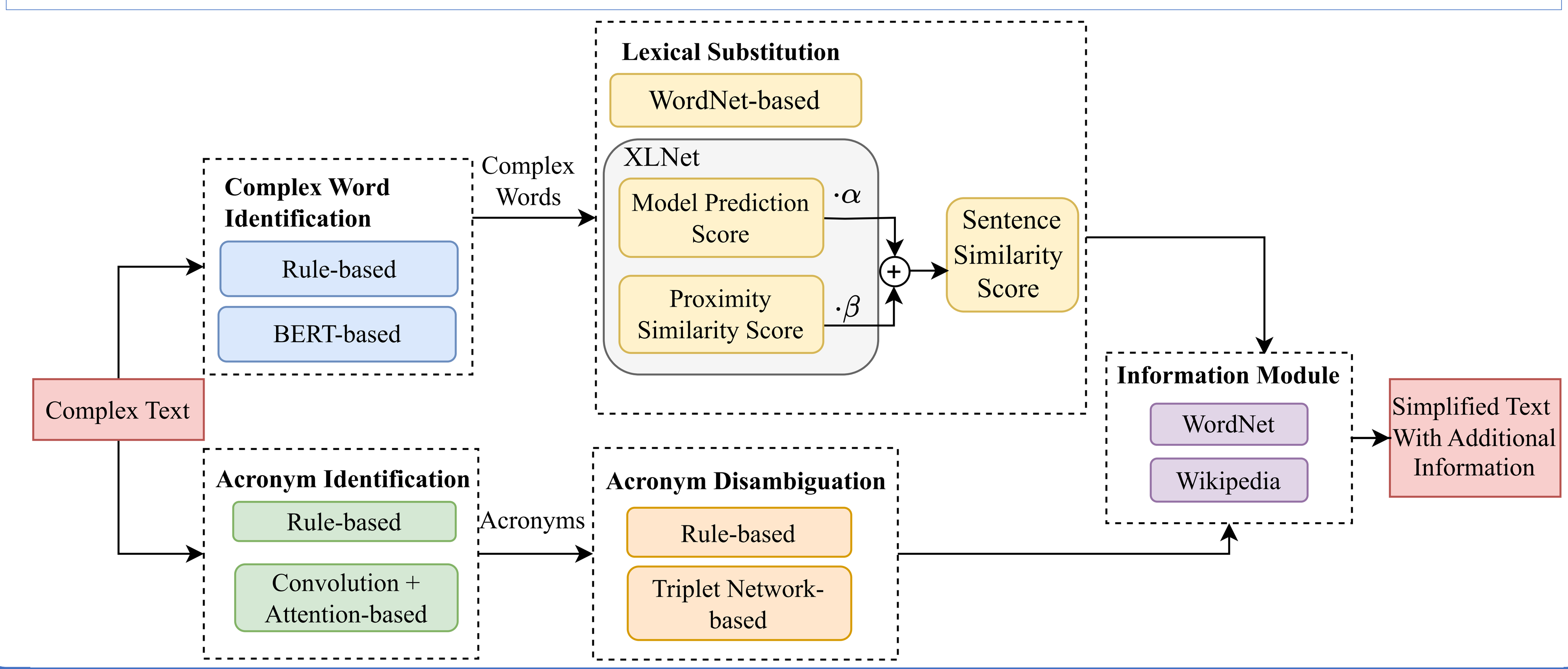
- Defined anchor, positive, and negative sentences.
- Frequency-based, BERT-based methods.

Method	F1 (%)	
	SDU	MeDAL
BERT-based	59.73	44.39
XLNet+embs	84.24	74.91
Triplet-Network-based	**85.70**	**75.19**

Information Module

- 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.

Proposed Framework



Comparison

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.

References

1] Shardlow, M. (2014). A survey of automated text simplification. International Journal of Advanced Computer Science and Applications, 4, 1 (2014), 58–7.
[2] McCarthy, D., & Navigli, R. (2009). The English lexical substitution task. Language resources and evaluation, 43, 2 (2009), 139–159.
[3] Seneviratne,. Daskalaki, E.; Lenskiy, A.; and Suominen, H., 2022b. m-networks: Adapting the triplet networks for acronym disambiguation. In Proceedings of the 4th Clinical Natural Language Processing Workshop, 21–29. Association for Computational Linguistics, Seattle, WA.
[4] Seneviratne, S Daskalaki, E.; Lenskiy, A.; and Suominen, H., 2022a. Cilex: An investigation of context information for lexical substitution methods. In Proceedings of the 29th International Conference on Computational Linguistics, 4124–4135.

Acknowledgement

This research was funded by and has been delivered in partnership with Our Health in Our Hands (OHIOH), a strategic initiative of the Australian National University (ANU). We gratefully acknowledge the funding from the ANU School of Computing and OHIOH for the first author's PhD studies. 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.