SENTA: Sentence Simplification System for Slovene

Aleš Žagar[1], Matej Klemen[1], Iztok Kosem[1,2,3], Marko Robnik-Šikonja[1]
[1] University of Ljubljana, Faculty of Computer and Information Science,
Večna pot 113, 1000 Ljubljana, Slovenia
[2] University of Ljubljana, Faculty of Arts, Aškerčeva cesta 2, 1000 Ljubljana, Slovenia
[3] Jozef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
{ales.zagar, matej.klemen, marko.robnik}@fri.uni-lj.si
iztok.kosem@ijs.si

Abstract

Ensuring universal access to written content, regardless of users' language proficiency and cognitive abilities, is of paramount importance. Sentence simplification, which involves converting complex sentences into more accessible forms while preserving their meaning, plays a crucial role in enhancing text accessibility. This paper introduces SENTA, a system for sentence simplification in Slovene. The system consists of two components. First, a neural classifier identifies sentences that require simplification, and second, a large Slovene language model based on T5 architecture is fine-tuned to transform complex texts into a simpler form, achieving an excellent SARI score of 41. Both automatic and qualitative evaluations provide important insights into the problem, highlighting areas for future research in multilingual applications, and fluency maintenance. Finally, SENTA is integrated into a freely accessible, user-friendly user interface, offering a valuable service to less-fluent Slovene users.

Keywords: Sentence Simplification, Text Accessibility, Large Language Models

1. Introduction

With an abundance of text-based content in our daily lives, the ability to understand and access information is a fundamental need of individuals, regardless of their age, education level, and linguistic abilities. However, the complexity of language often acts as a barrier, preventing many from fully engaging with written material. Text simplification approaches can provide a valuable aid and are therefore highly relevant research and development topics.

Sentence simplification transforms complex sentences into more accessible and understandable forms while preserving the essential meaning and context. The task plays a pivotal role in enhancing accessibility, especially for individuals with cognitive disabilities, non-native speakers, and those with limited literacy skills. Moreover, sentence simplification is a useful component in natural language processing, language generation, and automated content summarization.

Our goal is to develop a Slovene sentence simplification system primarily for users with a cognitive impairment, but also for other users aiming to make their texts more comprehensible. In this paper, we present the simplification methodology and its initial evaluation, while an end-user study will be conducted in further work. Our approach consists of two main steps, outlined in Figure 1. In the first step, we use a classifier to determine whether a given sentence is already simple enough or has to be further simplified. In the second step, we fine-tune a large language model (LM) based on T5 transformer architecture (Raffel et al., 2020) on an automatically translated sentence simplification dataset to generate simplified sentences. The results show that our best model outperforms a 200x larger LM. Our contributions are:

1. State-of-the-art Slovene sentence simplification system.

2. Extensive quantitative and qualitative evaluation of the results.

We structured the paper into further five sections. In Section 2, we outline related works, in Section 3, we present the used data, and, in Section 4, the used methodology. We discuss the evaluation and

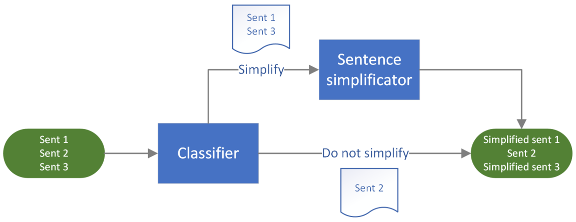

Figure 1: Our sentence simplification system. Classifier estimates which sentences need a simplification and which do not. The sentence simplifier generates simplified versions of the complex sentences. The final document consists of a mixture of simplified and original sentences (already simple in their original form).

LREC-COLING 2024, pages 14687–14692
20-25 May, 2024.

the obtained quantitative and qualitative results in Section 5. We present the main conclusions and possible further work in Section 6.

2. Related Work

In this section, we provide an overview of some of the notable datasets and methods used in sentence simplification research.

The Simple PPDB dataset (Pavlick and Callison-Burch, 2016) is a valuable resource for sentence simplification tasks. It is derived from the larger PPDB dataset and contains a list of 4.5 million simplifying paraphrase rules. WikiLarge (Zhang and Lapata, 2017) contains approximately 300k sentence pairs aggregated from 3 datasets (Kauchak, 2013; Woodsend and Lapata, 2011; Zhu et al., 2010). TurkCorpus (Xu et al., 2016) is a crowd-sourced dataset that focuses on English sentence simplification. It consists of 2,359 original sentences from English Wikipedia, each with eight manual reference simplifications. The Newsela (Xu et al., 2015) dataset aligns news articles of different reading levels. It offers multiple versions of the same news article, each tailored to a specific reading comprehension level.

Many sentence simplification approaches use supervised learning and deep learning models, mostly sequence-to-sequence neural networks, on annotated datasets of complex and simplified sentences (Xu et al., 2016; Sheang and Saggion, 2021).

Unsupervised and semi-supervised approaches aim to simplify sentences without relying on large amounts of labeled training data. Dehghan et al. (2022) propose a combination of text generation and text revision. It uses an iterative framework that incorporates explicit edit operations, including paraphrasing, to improve control and interpretability.

While much of the research focuses on English, sentence simplification can be a multilingual task. MUSS (Martin et al., 2022) experimented with their approach on English, French, and Spanish. It uses sentence-level paraphrase data obtained by mining paraphrases from Common Crawl in combination with controllable generation mechanisms to adjust the length and lexical complexity of a simplified candidate.

To assess the quality of simplified sentences, various evaluation metrics have been proposed. These include BLEU, SARI, and Flesch-Kincaid readability scores. In this work, we use SARI (Xu et al., 2016) which assesses the quality of words added, deleted, and retained by the systems by comparing it to both references and the input sentence.

3. Dataset

The Newsela English dataset (Xu et al., 2015) is a foundational resource for assessing the effectiveness of sentence simplification techniques. It offers a diverse collection of news articles, each curated to offer multiple versions of the same article at varying reading levels, ranging from elementary to advanced. This structure allows readers to access content that aligns with their comprehension skills, while the diversity of texts enables researchers to investigate the complexities of sentence structures, vocabulary, and content while facilitating the evaluation of sentence simplification approaches. We employed the DeepL translation system [1] to automatically translate the Newsela articles from English to Slovene. We used the sentence-aligned version of the dataset.

Recognizing the issues of automatic translation, we also created an evaluation dataset of 1583 authentic sentences from a wide range of sources and genres, ensuring comprehensive coverage of the Slovene language (Kosem et al., 2022a; Žagar et al., 2022; Kosem et al., 2022b).

3.1. Data preprocessing

Starting with the machine-translated dataset, we first manually fixed minor formatting issues where multiple examples were incorrectly grouped, possibly due to imperfect tokenization or sentence alignment. As we were interested in producing as simple sentences as possible, we filtered the dataset, keeping only examples with the easiest comprehensibility level (i.e., target grade "V4"). To steer models away from copying the input without making any modifications, we tokenized the complex and simple sentences and discarded examples in which the intersection over the union of lower-cased tokens is above 0.7[2]. Finally, to produce a split into training, validation, and test set, we grouped the examples by their document identifier, and took examples from 125 documents for the test set, from 75 for the validation set, and the rest for the training set. In total, we obtained $35,455$ training, $2,849$ validation, and $4,818$ test examples.

On this data, we trained two models. For the first model in the processing pipeline that classifies if a sentence requires simplification, we produced the binary label by marking all the complex (source) sentences as "requiring simplification" (positive class) and all the simple (target) sentences as "not requiring simplification" (negative class). The second model that generates simplified sentences uses the unchanged dataset, using complex

[1] https://www.deepl.com/translator

[2] The number was determined manually on a small sample of training examples.

sentences as the input and simple sentences as the target.

4. Sentence Simplifiers

In this section, we present the methodology employed in our study, outlined in Figure 1. While testing several variants. our final approach leverages an initial SloBERTa-based (Ulčar and Robnik-Šikonja, 2021) complex sentence detector, and SloT5-based (Ulčar and Robnik-Šikonja, 2023) sentence simplifier. We describe both components of our approach, together with technical details such as the tokenizer parameters for the classification model and the generator, as well as the training and inference configuration.

4.1. Complex sentence detector

As not all input text necessarily needs to be simplified, we built a classification model that identifies whether an input sentence is already comprehensible enough, and thus does not need further simplification. Note that the next component in the pipeline, i.e., the simplifier, is trained to map complex sentences to simple ones and is only trained on pairs of complex and simple sentences. During training, the simplifier does not see any simple sentences on its input, therefore it makes sense to remove such sentences also during its application to reduce the potential out-of-sample bias.

We approach the problem of complex sentence detection as a binary text classification task, i.e., we determine whether a sentence is simple or not. For that purpose, we use the dataset described in Section 3.1 to fine-tune the pre-trained Slovene masked LM SloBERTa (Ulčar and Robnik-Šikonja, 2021), based on RoBERTa (Liu et al., 2019) architecture. To reduce memory usage and enable batch processing, we truncated the inputs to the maximum length of 65, determined as the 99th percentile of input lengths in the training dataset.

4.2. Sentence simplifier

For sentence simplification, we tested several approaches: Slovene SloT5 (small and large variants), multilingual mT5, and GPT3.5-turbo (ChatGPT).

We fine-tuned two variants of SloT5 (Ulčar and Robnik-Šikonja, 2023), small and large, both pretrained on the Slovene language and based on the T5 architecture (Raffel et al., 2020). The small SloT5 model, i.e., T5-sl-small on HuggingFace[3] contains 60 million parameters and is the smallest of our models. As the final models will be included in a web application, its compact size is an advantage. The SloT5-large model[4] is significantly larger with 750 million parameters.

In contrast to the language-specific SloT5 models, the mT5 model, i.e mT5-small[5], is designed to be multilingual, supporting various languages. With 300 million parameters, it offers a broad language coverage.

Our last model is GPT3.5-turbo[6], a powerful, large-scale, multilingual model with presumed 175 billion parameters. It offers a broad range of language support and versatility.

When the above complex sentence classifier identifies a sentence in need of simplification, we forward it to the simplifier. While we tested several models and their parameters, the fine-tuning parameters for our top-performing model were set as follows: the number of training epochs was established as 3, and the learning rate was set to 5e-5.

For the text simplification model, we used a tokenizer with the following parameters: *max_length* was set to 128, *truncation* to True, and *padding* to "max_length". At the inference step, we employed beam-search multinomial sampling with the following parameters: *maximum length* of 128, *beam_size* of 5,[7] *sampling* was enabled, *top-k* value was set to 5, and *temperature* to 0.7.

5. Results

In this section, we present automatic and manual evaluation of the results.

5.1. Complex sentence detector

The complex sentence detector achieves a classification accuracy of 0.80; the detailed results achieved by the complex sentence detector are presented in Table 1. The metrics are all relatively stable at around 0.80 (0.78 - 0.82); for the positive class, the precision is slightly higher than the recall, and for the negative class, the reverse is true.

5.2. Automatic evaluation

The results of the quantitative evaluation are based on the SARI score and presented in Table 2. The

[3] https://huggingface.co/cjvt/t5-sl-small

[4] https://huggingface.co/cjvt/t5-sl-large

[5] https://huggingface.co/t5-small

[6] gpt-3.5-turbo-0613: https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates

[7] A beam size of 5 is a balance between generating high-quality outputs and maintaining reasonable computational efficiency. A beam size of 1 equals greedy search and in our case did not produce satisfactory results. Larger beam sizes did not produce better results; they were on par with our final beam size. Therefore, we used 5 mainly because of better computational efficiency.

Table 1: Classification Metrics

Class	Precision	Recall	F1
Positive	0.82	0.78	0.80
Negative	0.79	0.83	0.81

Table 2: Results. The size of the GPT3.5-turbo model is marked with ? as its size is not publicly disclosed and the estimate of 175B parameters is made based on its predecessor.

Model	Language	# params	SARI
SloT5-small	Slovene	60M	39.79
mT5-small	multiling.	300M	39.09
SloT5-large	Slovene	750M	**41.01**
GPT3.5-turbo	multiling.	175B?	38.76

Table 3: Summary of SloT5-large manual evaluation, showing numbers of sentences (#S) with different properties. Categories under problematic sentences overlap.

Category	**#S**
Total Examples	1583
Non-Complex Sentences	396
Satisfactory Simplified Sentences	611
No Simplification Required	11
Identical Sentences	-
Problematic Sentences	561
- Incorrect Information due to Simplification	198
- Lack of Sufficient Information	309
- Logically Coherent Simplified Sentences	401
- Incoherent Simplified Sentences	153
- Simplified Sentences with Syntactical Errors	53

SloT5-small model achieved the SARI score of 39.79, which suggests that it performs reasonably well in terms of semantic adequacy, relevance, and readability improvement for Slovene text data. The SloT5-large model, also tailored for Slovene, stands out with a notable SARI score of 41.01. It achieves this impressive result with a substantially larger size of 750 million parameters. The higher SARI score suggests enhanced performance in terms of semantic adequacy, relevance, and readability improvement compared to the smaller SloT5-small model.

Finally, the GPT3.5-turbo achieved the SARI score of 38.76, which is slightly less than all the T5 models. This result might indicate that, despite its size and multilingual capabilities, the T5 models can be better tuned in our specific simplification context and Slovene language.

5.3. Manual evaluation

For our best model, SloT5-large, we also performed a manual internal evaluation on the evaluation dataset with the help of an expert linguist. Table 3 shows the summary of the human evaluation focused on sentence simplification.

A more detailed analysis reveals that simplifications often remove or significantly shorten subordinate clauses. This works well when a subordinate clause contains non-essential information. The position of this non-essential clause in the sentence does not affect the quality of simplification. On the other hand, if a subordinate clause contains essential information in the sentence, its omission or oversimplification often results in poor results.

Examples of good simplifications:

• Napad je bil povod, da so ZDA vstopile v drugo veliko vojno, s čimer je konflikt, ki je bil dotlej bolj ali manj evropskega značaja, postal svetoven. (eng. *The attack was the reason the US entered the second large war, making the conflict, which had been more or less European till then, world-wide.*)

○ Napad je bil povod, da so ZDA vstopile v vojno. (eng. *The attack was the reason the US entered the war.*)

• Zelene trate botaničnega vrta so že pobeljene, sicer pa se bodo zvončki v prihodnjih dneh še razcveteli. (eng. *The green lawns of the botanical garden are already whitened, but the bluebells will bloom in the coming days.*

○ Zelene trate botaničnega vrta so že pobeljene. (eng. *The green lawns of the botanical garden are already whitened.*

• Odkritje je navdušeno spremljalo tudi naše časopisje, pri čemer je šlo predvsem za povzemanje tujih člankov. (engl. *The discovery was also enthusiastically followed by our newspapers, with the majority summarizing foreign articles.*

○ Odkritje je navduševalo tudi naše časopise. (eng. *Our newspapers were also fascinated by the discovery.)*

Examples of bad simplifications:

• Sopotnika sta takoj poklicala reševalce, vendar je bilo že prepozno, da bi mu lahko pomagali. (engl. *The two fellow passengers immediately called the ambulance, but it was too late to help him.*

○ Sopotnika sta takoj poklicala reševalce. (eng. *The two fellow passengers immediately called the ambulance.*

• Novomeški policisti so v zadnjih dneh obravnavali voznika kombija, ki je po avtocesti vozil vzvratno ter nato parkiral ob vozišču na travi. (eng. *In the last few days, the Novo mesto police officers dealt with the driver of the van, who was driving in reverse on the highway and then parked on the grass next to the carriageway.*)

○ Novomeški policisti so v zadnjih dneh obravnavali voznika kombija. (eng. *In the last few days, the Novo mesto police officers dealt with the driver of the van.*)

In the manual evaluation, the main focus was on whether the main information of the sentence was retained without making the new sentence inaccurate or not informative enough. Moreover, the aim of simplification was also to shorten the sentences, which was also reflected in training data. As the provided examples of good simplifications show, sometimes the additional information, provided as a subordinate clause, was omitted. In addition, the omitted or shortened information was also considered from the perspective of difficulty. For example, in the sentence about the US and world war, one could argue that the information about the conflict becoming worldwide should be retained, however the stress is on the US entering the war, and the omitted part contains some difficult vocabulary in Slovene, e.g., *značaj*, *dotlej*.

Table 4: Prompt for the GPT3.5 model.

Task: Simplify the following sentences in Slovene while maintaining naturalness and readability for cognitively impaired people.
Original Sentence: Genetic engineering has expanded the genes available to breeders to utilize in creating desired germlines for new crops.
Simplified Sentence: New plants were created with genetic engineering.
Original Sentence: A naval mine is a self-contained explosive device placed in water to destroy ships or submarines.
Simplified Sentence: A naval mine is a bomb placed in water to destroy ships or submarines.
Original Sentence: A boot is a type of footwear that covers at least the foot and the ankle and sometimes extends up to the knee or even the hip.
Simplified Sentence: A boot is a type of footwear that protects the foot and ankle.
Guidelines:
- **Language**: Provide your response in Slovene.
- **Explain Complexity**: If you encounter complex words, provide a simpler explanation or synonym.
- **Complex Word Removal**: Remove any words too complex for the target audience.
- **Sentence Splitting**: Break complex sentences into simpler ones, not exceeding 10 words.
- **Conciseness**: Keep sentences short and clear.
Original Sentence: SLOVENE SENTENCE

One particular problem, which is evident from the above examples but is not limited to sub-ordinate clauses, could be called the Rheme problem, i.e., the cases when essential information is provided at the end of the sentence, especially a longer one. This is problematic because it can be often observed that the final parts of the sentence are simplified or even omitted.

One aspect where the simplification performs fairly well is the replacement of individual words with their (more general and frequent) (near)-synonyms, e.g., zvrstiti se –> potekati (eng. *to take place*), ugnati –> premagati (eng. *to beat*), poudariti –> praviti (eng. *stress* –> *say*).

6. Conclusion

We present the first Slovene sentence simplification system SENTA[8]. The system represents a useful solution for the Slovene language and significantly enhances text accessibility in the Slovene language. By seamlessly integrating a classifier for efficient sentence selection and a SloT5-based simplifier, the SENTA system offers a practical solution for addressing language complexity and improving text comprehension.

The system's automatic and manual evaluations demonstrate its proficiency in generating simplified sentences, with the SloT5-large model being the best with a SARI score of 41. These results indicate that SENTA will effectively enhance text accessibility for diverse user groups, including those with cognitive disabilities, non-native speakers, and individuals with limited literacy skills.

The analysis of the results also points to many further challenges in sentence simplification, including handling idiomatic expressions, ensuring readability, and conveying relevant information from subordinate clauses. Furthermore, future research directions shall explore the integration of domain-specific knowledge, address potential biases in simplification, and enhance the adaptability of models to varying reading levels.

Acknowledgements

The work was partially supported by the Slovenian Research Agency (ARRS) core research programme P6-0411, as well as projects J7-3159, CRP V5-2297, and L2-50070. The work was also supported by the SLOKIT project (Upgrading CLARIN.SI), funded by the Ministry of Culture.

7. Bibliographical References

Mohammad Dehghan, Dhruv Kumar, and Lukasz Golab. 2022. Grs: Combining generation and revision in unsupervised sentence simplification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 949–960.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1537–1546.

Iztok Kosem, Jaka Čibej, Kaja Dobrovoljc, Tomaž Erjavec, Nikola Ljubešić, Primož Ponikvar, Mihael Šinkec, and Simon Krek. 2022a. Monitor corpus of slovene trendi 2022-10. Slovenian language resource repository CLARIN.SI.

Iztok Kosem, Eva Pori, Aleš Žagar, and Špela Arhar Holdt. 2022b. Corpus of slovenian textbooks ccUčbeniki 1.0. Slovenian language resource repository CLARIN.SI.

[8] SENTA is also a text analysis system. Link to the complete web application: https://senta.cjvt.si/. Link to repositories: https://github.com/clarinsi?q=senta&type=all&language=&sort=

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*.

Louis Martin, Angela Fan, Éric Villemonte De La Clergerie, Antoine Bordes, and Benoît Sagot. 2022. Muss: Multilingual unsupervised sentence simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664.

Ellie Pavlick and Chris Callison-Burch. 2016. Simple ppdb: A paraphrase database for simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 143–148.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352.

Matej Ulčar and Marko Robnik-Šikonja. 2021. SloBERTa: Slovene monolingual large pretrained masked language model. In *Proceedings of the 24th International Multiconference – IS2021 (SiKDD)*.

Matej Ulčar and Marko Robnik-Šikonja. 2023. Sequence-to-sequence pretraining for a less-resourced Slovenian language. *Frontiers in Artificial Intelligence*, 6:932519.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Aleš Žagar, Matic Kavaš, Marko Robnik-Šikonja, Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Marko Ferme, Mladen Borovič, Borko Bošković, Milan Ojsteršek, and Goran Hrovat. 2022. Corpus of academic slovene KAS 2.0. Slovenian language resource repository CLARIN.SI.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *EMNLP 2017: Conference on Empirical Methods in Natural Language Processing*, pages 584–594. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361.