University *of Ljubljana*
Faculty *of Computer and Information Science*

Reviving Historical Slovene: Exploring Large Language Models for Translating Modern Slovene Texts to the Past

Martin Jurkovič, Blaž Pridgar, and Jure Savnik

Abstract

This report presents a study on paraphrasing sentences from modern Slovene to old Slovene using generative large language models. The focus of the study is on fine-tuning multiple Slovene text models on a corpus of old Slovene texts from the IMP corpus. The IMP corpus, which includes slovene texts from the 16th to the 20th century, provides a valuable resource for training and evaluating the models. The study compares the performance of different models, including Slovene T5 small, Slovene T5 large, and multilingual T5, using the ROUGE metric for evaluation. The results show that the fine-tuned models achieved high accuracy, with scores above 90% on average. The findings suggest that generative text models can effectively paraphrase modern Slovene sentences into old Slovene, opening up possibilities for language translation and historical text analysis.

Keywords

Public domain book corpus, Paraphrasing, Large Language Models, Historical Slovene literature

Advisors: Slavko Žitnik

Introduction

With the public demo of ChatGPT the field of natural language processing (NLP) has seen the most attention from the public eye any machine learning (ML) field has ever seen. With insurmountable amount of written language, be it books, articles, blogs, tweets, many of the sources can't be publicly used in the ML domain because of copyright issues. That's why we decided to focus on books which are in public domain, meaning they are not being subject to copyright or other legal restrictions. One of the most popular public domain book repositories is Project Gutenberg [1] (*www.gutenberg.org*), which holds only one Slovenian book. We decided to use a IMP corpus by Erjavec T. (2015) [2] of old Slovene texts.

Given the prevalence of old Slovene language within the vast collection of books available in the public domain, we made a deliberate choice to center our efforts on the task of paraphrasing sentences from modern Slovene to historical Slovene. To accomplish this, we employed a fine-tuning approach with multiple generative Slovene text models, namely Slovene T5 small, Slovene T5 large, and multilingual T5. This allowed us to conduct a comprehensive performance evaluation, comparing the effectiveness and fidelity of each model in paraphrasing old Slovene texts.

Related work

Public domain book corpus

Gerlach, M. and Font-Clos, F. built and processed a *Standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics* [1]. They processed the books by taking raw text, extracting only text, tokenizing this text with a tokenizer from the *nltk* library and produced word counts. Besides the title and author of the book they have also provided a broad characterization of the books in terms of their length, language, date of publication and subject. They have also quantified variability in the corpus based on subject and time of publication.

Another notable paper is *Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books* by Zhu, Y., Kiros, R. [3]. They created a corpus of 11.038 books from the web and provided six summary statistics. The researches downloaded every free book longer than 20.000 words from the website *www.smashwords.com*, which contains e-books by independent authors. They have also aligned the books with corresponding movie releases and mostly focus on that topic, with almost no mention of corpus preprocessing or summarization.

Generative text modelling

A group of four researches from Huawei Noah's Ark Lab wrote a paper on fine tuning a generative text model for generating Chinese poetry [4]. This is a similar task to what we aspire to achieve with Slovene literature. In their example they fine tuned a model trained on a corpus of Chinese news with 235M sentences with an additional dataset of around a million Chinese poems.

Another relevant paper written by Jieh-Sheng and LeeJieh Hsiang describes how they fine tuned OpenAIs GPT-2 with a corpus of more than 500.000 patent claims to develop a pre-trained model for generating patent claims [5].

Large Language Models

Our study was focused on fine tuning a Large Language Model (LLM) for a specific task, so we relied heavily on the work of others with the use of a range of open source pre-trained LLMs. Since our task is Slovene based we started with the Slovene T5 and Slovene GPT models, but the GPT based model soon proved to be unsuitable for the task at hand because of problems with fine tuning. M. Ulčar and M. Robnik-Šikonja presented a small and large Slovene T5 models with 60 million and 750 million parameters respectively in their paper *Sequence to sequence pre-training for a less-resourced Slovenian language* [6]. They also compared the performance of these models to other LLMs in several tasks and found that the small Slovene model outperform the small multilingual T5 (mT5) published by a group of Google researcher in their paper *mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer* [7]. Interestingly it was opposite for the large Slovene model and large mT5. Ulčar and Šikonja explained this difference in performance with the lack of Slovene training data for the large model to sufficiently train all its parameters.

Evaluation Metric

ROUGE (or Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics first introduced by Chin-Yew Lin in his paper *ROUGE: A Package for Automatic Evaluation of Summaries* for evaluating models on summary tasks. While ROUGE was designed specifically for the evaluation of summarization tasks, it can also be used to evaluate the quality of machine translation or paraphrase systems. ROUGE calculates the overlap between the machine-generated output and the reference text using various measures - ROUGE-1 (measures the overlap between unigrams), ROUGE-2 (measures the overlap between bigrams), ROUGE-L: (measures the longest common sub-sequence which takes into account word order).

Methods

Corpus analysis & preprocessing

We used a well composed corpus of texts in historical Slovene IMP corpus by Erjavec T. (2015) [2]. Most texts are from the 19th century, although older books are also included, with the oldest book being the Dalmatin's bible from 1584. The corpus includes the glossary, which shows archaic spellings of words and their glosses in contemporary Slovene. The concodardancer supports the use of corpus-based methods for diachronic studies of the Slovene Language. We use the integral resources encoded in XML which are useful for the development of Human Language Technologies for processing historical Slovene.

The digital library contains over 650 units (books, newspapers and manuscripts) from the end of the 16th century to 1918 with the majority from 1850 on-wards. Each document contains the title, author, year and document ID. The corpus contains books from the following sources:

- **WIKI**: Texts from the Wikisource project *Slovene literary classics* [8]
- **FPG**: the AHLib [9] collection of books translated from German into Slovene (1848 - 1918)
- **NUK**: older books (1750-1820) and selected issues of one newspaper (1850-1900) prepared in the scope of the EU IMPACT project by the National and University Library of Slovenia (NUK)
- **ZRC**: small samples of three religious texts (1584, 1695, 1784) prepared by the Scientific Research Center of the Slovene Academy of Sciences and Arts.

The 650 units consist of facsimiles with 45,000 pages as well as hand-corrected and structured transcriptions. The hand annotated corpus has 300,000 tokens, where each word is tagged with it's modernised word form, lemma, part-of-speech and in case of archaic words it's nearest contemporary equivalents. The complete set of IMP resources is encoded following the Text Encoding Initiative Guidelines TEI P5, a parametrisable annotation vocabulary and encoding scheme, with which it is possible to construct project-specific XML schemas and use them to validate and document the resource encoding.

We have processed the TEI document and first for each word in the texts extracted their corresponding modern word form and the lemma. Then we have converted the words into sentences and got a little less than one million sentences. However we noticed that most of the modern forms of words are not translated at all, but just contain the same old word repeated. So we removed all of the sentences which do not have a proper modern translation and were left with $380,658$ sentences, which corresponds to $637,526$ unique tokens.

Fine tuning

For the first part of our study we focused on fine tuning the Slovene T5 model. Since the paper by Ulčar and Šikonja suggested that the small model might perform better in some tasks we fine tuned both the small and the large model so we could make a comparison. The fine tuning process consists of the following steps:

Data Preparation is needed for data tokenization and data collation. First we transform the input sentences into lists of tokens and then we utilize a data collator to deal with the padding for dynamic batching.

Choice of Metric is dependent on the task at hand and is not completely necessary, but is useful to get feedback during model training. We chose to use the ROUGE metric.

Setting the Training Parameters is the only part of the fine tuning where we have the power to influence the process. The important hyper-parameters are learning rate, number of epochs, batch size, and weight decay but additional parameters can also be added. An important observation was that the Slovene T5 models don't work with the *fp16* parameter turned on, which is meant to speed up the training process by using mixed precision (not all variables need to be stored in full 32-bit floating point precision).

Results

We tested the models we fined tuned for old Slovene on 5000 test cases. For evaluation we used ROGUE metric described above. We compared our models to the baseline; which is the ROUGE score obtained by comparing sentences in modern Slovene with the ones in the old Slovene. Results can be seen in Table 1. One thing we have to take into consideration when comparing these two scores is that not all sentences are from the same time period. Some are older and some are from more modern Slovene. So in some cases the sentence we translated was already pretty close to the one which we then compared it to. In those cases our model performed worse than sentence that was not translated. That is why we also wanted too see how well our model performs, if we only evaluate it on the words that are different in modern and old Slovene sentences. Results for this can be seen in Table 2. Lastly in Table 3, we can observe the percent of words that were translated and the model also translated, and number of words the model unnecessarily added. We also tested how different values of the two parameters (top-k: The number of highest probability vocabulary tokens to keep for top-k-filtering; top-p: only the smallest set of most probable tokens with probabilities that add up to top-p or higher are kept for generation) affect translations.

Overall

All the fine tuned models performed really well with the average accuracy of more than 90%, which was even higher when we used different parameters. We could improve these results even further by making a more formal search for optimal parameters. Even when checking Rouge scores only on words that were changed presented in Table 2, we can see that the model translated words really well, since it raised all the scores for around 80% in comparison to untranslated words.

model	Rouge1	Rouge2	RougeL	RougeLsum
No translation	0.80	0.68	0.80	0.80
T5-small	0.92	0.86	0.92	0.92
T5-small parameters	0.95	0.91	0.95	0.95
mT5-base	0.93	0.87	0.93	0.93
mT5-base parameters	0.95	0.91	0.95	0.95
T5-large	0.91	0.86	0.91	0.91
T5-large parameters	0.94	0.90	0.94	0.94

Table 1. Table shows the Rouge1, Rouge2, RougeL and RougeLsum scores for evaluating text that had no translation and for all the models we fine tuned and tested.

model	Rouge1	Rouge2	RougeL	RougeLsum
No translation	0.07	0.01	0.07	0.07
T5-small	0.87	0.44	0.087	0.87
T5-small parameters	0.92	0.48	0.92	0.92
mT5-base	0.89	0.45	0.89	0.89
mT5-base parameters	0.92	0.48	0.92	0.92
T5-large	0.83	0.41	0.83	0.83
T5-large parameters	0.89	0.47	0.89	0.89

Table 2. Table shows the Rouge1, Rouge2, RougeL and RougeLsum scores for texts that had no translation and for all the models we fine tuned and tested. When generating translated text top-k was set to 100 and top-p to 0.6.

The only exception are Rogue-2 scores which are lower because translated words sometimes had additional or removed letters and because of words our model inserted mistakenly. We also have to take into consideration that the old Slovene sentences contained an old Slovene "s" which no model translated to it, but just replaces it with modern "s", which also affected the scores a bit. Two examples of how models performed can be seen in Table 4.

Model	Words Changed [%]	Added Words
T5-small	100	1374
T5-small parameters	100	1110
mT5-base	100	476
mT5-base parameters	100	211
T5-large	99	2258
T5-large parameters	100	1950

Table 3. Table shows percent of words that had translation in corpus that our model actually translated and number of words it unnecessarily added to sentences.

Parameters

Finding optimal parameters for machine translation is very important for making the model more stable with much better performance. For example when we try to translate "Ti hočeš Tonček" to "Ti ozhesk Tonzhek" with T5-large model without parameters we get a lot of different results. Some of the examples are "Ti možmash Tonzhek.", "Ti ozhesh Tonzek.", "Ti zhlovesh Bentushek.", "Ti želioš Tonzhek". Most of these results are not really good, but amidst them is the almost right translation. This means that the quality of our results can be really different.

Modern:	in zašribaj te besede nanjo?
Historical:	noi sashribei te besiede nanjo?
T5-large:	Naras shibe ti desesied nanjo?
mT5-base:	noi sashribei te besiedé nanjo?
T5-small:	Noi sashribei te Besiede na?
T5-large (parameters):	Na sashribei te besiede nanjo?
mT5-base (parameters):	noi sashribei te besiede nanjo?
T5-small (parameters):	Noi sashribei te besiede na?
Modern:	priteče vaša gnada, godci so že tukaj; mladost od cele vasi, fantje, in dekliči, zunaj čakajo, in prosijo baron.
Historical:	pertezhe Vasha Gnada, godzi so vshe tukej; mladost od zele vassy, fantje, inu deklizhi, sunej zhakajo, inu prossijo Baron.
T5-large:	Pertezhe Vasha Gnada, godzi so vshe tukej; mladost od zele vassy, fantje, inu deklizhi, sunej zhakajo, inu prossijo Baron.
mT5-base:	Pertezazhé Vasha Gnada, gôzhi so vshe tukej; mladost od zele vassy, fantje, inu deklizhi, sunej zhakajo, inu prossijo Baron.
T5-small:	Pertezhhe Vasha Gnada, godzi so vshe tukej; mladost od zele vassha, fantje, inu deklizhhi, sunej zhakajo, inu prossijo Baron.
T5-large (parameters):	Pertezhe Vasha Gnada, godzi so vshe tukej; mladost od zele vassy, fantje, inu deklizhi, sunej zhakajo, inu prossijo Baron.
mT5-base (parameters):	Pertezhe Vasha Gnada, godzi so vshe tukej; mladost od zele vassy, fantje, inu deklizhi, sunej zhakajo, inu prossijo Baron.
T5-small (parameters):	Pertezhe Vasha Gnada, godzi so vshe tukej; mladost od zele vassy, fantje, inu deklizhi, sunej zhakajo, inu prossijo Baron.

Table 4. Table shows sentences in modern Slovene, old Slovene, the ones translated with T5-Large, mT5-base and T5-small. Below there are examples of translations of all models where parameters were changed for a better performance.

But changing top-k and top-p parameter makes transitions really stable. Setting them to $top-k = 100$ and $top-p = 0.6$ we consistently get result that start with "Ti ozhesh" and ends with "Tonzhek" or something similar.

mT5-base

In general, it is our considered opinion that this particular model exhibited superior performance across various metrics, including translation accuracy, word count, and lexical modifications. However, it is worth noting that in certain instances, the model received a lower score due to alterations in word plurality, the introduction of acute accents, and occasional additions or omissions of letters within words. Nevertheless, this fine-tuned model demonstrated commendable overall efficacy.

T5-large

This model demonstrated a notably inferior performance when it came to translating texts into Old Slovene. Multiple instances revealed inadequate word translations, as seen in the following examples: "električna $\rightarrow$ eglična", "milančana $\rightarrow$ miles-čana", and "ljubezni $\rightarrow$ Ivan". We can also so an inadequate performance for the first example in table 4. Furthermore, this model exhibited similar issues to the other two models, including the gratuitous addition or removal of letters, superfluous acute accents, and the propensity for introducing unnecessary words. In terms of such linguistic deviations, this model surpassed the others in frequency.

T5-small

Similarly, this model demonstrated a commendable level of performance. While it encountered similar issues as the previous two models, its frequency of such issues was considerably lower compared to the T5-large model. This supports the hypothesis that there wasn't enough data to sufficiently train the bigger amount of parameters of the T5-large model. Notably, when compared to the mT5-base model, this model exhibited a greater propensity for introducing random word modifications and additional words. Nonetheless, the overall performance of this model appeared to be relatively comparable to that of the mT5-base model.

Discussion

Our study highlights the successful fine-tuning of language models for translating modern Slovene texts to historical Slovene. The results indicate that with appropriate parameter settings, the models achieve high translation accuracy. While the mT5-base model exhibited superior performance, the T5-small model also demonstrated commendable efficacy. The T5-large model, however, presented challenges in accurately translating texts to historical Slovene. Further research and optimization of parameters can potentially enhance the translation quality and make it more consistent.

In terms of future work, there are several directions that can be explored to further improve the translation of modern Slovene texts to historical Slovene. Firstly, it would be beneficial to experiment with additional Large Language Models (LLMs) to assess their performance and compare them to the models already tested. This can provide insights into the suitability of different LLM architectures for this specific translation task.

Moreover, a potential avenue for future research is to focus on translating texts to a more specific temporal period within historical Slovene. By narrowing down the time frame,

the models can be trained and fine-tuned on a more specific linguistic context, which may lead to more accurate and contextually appropriate translations. This approach can enhance the precision and historical accuracy of the translated texts, but the lack of training data could pose a problem.

Furthermore, exploring techniques for mitigating the issues encountered in the current study, such as erroneous word modifications or the addition of unnecessary words, remains a promising area for future investigation. This can involve implementing additional constraints, refining the fine-tuning process, or employing post-processing steps to ensure more faithful and linguistically sound translations.

By pursuing these avenues of future work, we can advance the field of translating modern Slovene texts to historical Slovene, leading to improved language models and contributing to the preservation and understanding of historical Slovene language and literature.

References

[1] Martin Gerlach and Francesc Font-Clos. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *CoRR*, abs/1812.08092, 2018.

[2] Tomaž Erjavec. The imp historical slovene language resources. *Language Resources and Evaluation*, 49, 01 2015.

[3] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR*, abs/1506.06724, 2015.

[4] Yi Liao, Yasheng Wang, Qun Liu, and Xin Jiang. Gpt-based generation for classical chinese poetry, 2019.

[5] Jieh-Sheng Lee and Jieh Hsiang. Patent claim generation by fine-tuning openai gpt-2. *World Patent Information*, 62:101983, 2020.

[6] Matej Ulčar and Marko Robnik-Šikonja. Sequence to sequence pretraining for a less-resourced slovenian language, 2023.

[7] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934, 2020.

[8] Wikivir: Slovenska leposlovna klasika. https://sl.wikisource.org/wiki/Wikivir:Slovenska_leposlovna_klasika. Accessed: 2023-04-30.

[9] Ahlib digital library. https://nl.ijs.si/ahlib/. Accessed: 2023-04-30.