FRI Natural language processing course 2021
Project report

University of Ljubljana
Faculty of Computer and
Information Science

Submission 2 - Initial implementation with results

Martin Jurkovič, Blaž Pridgar, and Jure Savnik

Abstract

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, may 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.

Since most of the books in public domain contain old Slovene language, we decided to focus on paraphrasing sentences from modern Slovene to old Slovene. We fine tuned multiple generative Slovene text models (Slovene T5, Slovene GPT2, multilingual T5) on this corpus and compared their performance on 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 quantifies 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 for 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-2 models. 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 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 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 model we fined tuned for old Slovene on almost 60000 test cases. For evaluation we used ROGUE metric described above. We compared our model to the baseline, which is the ROUGE score we obtained by scoring sentences in modern Slovene, the ones we than translated using our model, with the ones in the old Slovene. 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 made it worse and this is one of the reasons for bad score on some of the cases.

Table 1. Table shows the rouge1, rouge2, rougeL and rougeLsum scores for evalutaing text that had no translation and for all the models we fine tuned and tested.

model	Rouge1	Rouge2	RougeL	RougeLsum
No translation	0.8	0.68	0.80	0.8
T5-Large	0.70	0.61	0.70	0.70

T5-large

Seeing results in Table 1 it seems like fine tuning didn't help, since the scores went down. One of the reasons we already mentioned above, but the other reason why the results got worse is that translations are not always of the same quality. For example we will take sentence in modern Slovene ”Ti hočeš Tonček” which should translate to ”Ti ozhesh Tonzhek”. Some of the translations we get in this case are ”Ti možmash Tonzhek.”, ”Ti ozhesh Tonzek.”, ”Ti zhlovesh Bentushek.”, ”Ti želioš Tonzhek.” and so on. We can see that the results can be really different, some are really good while some are really bad. These means there can be a really big variability in the scores we get.

We got these results by translating using the default parameters for generating translations. Changing parameters when translating can be really beneficial. We tested what happens, if we change top_k (the number of highest probability vocabulary tokens to keep for top_k-filtering) and top_p (if set to < 1, only the smallest set of most probable tokens with probabilities that add up to top_p or higher are kept for generation), on the sentance mentioned above. By changing these two parameters we almost always got the result that started with ”Ti ozhesh” and ended or had some variation of name ”Tonček” at the end, but mostly ”Tonzhek”.

Discussion

The initial results of the first half of our study have not yielded desirable outcomes. Nonetheless, we have identified several factors that may have contributed to this outcome and developed a range of ideas to improve the performance of the models in the second part of the study. Firstly, we utilized default parameters to train the models, which we found to be a suitable starting point for fine-tuning. However, we recognize that there may be opportunities for improvement by exploring different learning rates to optimize the performance of the models. Therefore, we will conduct experiments to determine the most effective learning rate for achieving the best results.

Secondly, we recognize the importance of expanding our collection of base models available for fine-tuning. As noted by Ulčar and Šikonja, the large mT5 model has demonstrated superior performance compared to the large Slovene T5. Therefore, we will include the mT5 model as a candidate for fine-tuning in the second phase of the study. Additionally, we will experiment with the Slovene GPT-2 model to determine its efficacy in achieving the desired performance.

The final objective of our study is to achieve a comprehensive comparison of the performance of all the models under consideration and develop a more efficient Old Slovene text translator. Our findings will contribute to advancing the field of natural language processing and support efforts to preserve and promote the cultural heritage of Slovenia.

Acknowledgments

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.