FRI Natural language processing course 2023
Project report

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

Natural language processing course Latex Template

Aljaž Grdadolnik, Anže Mihevc, and Luka Galjot

Abstract

This paper discusses the use of deep learning models, particularly transformer-based language models like BERT, to implement sentence paraphrasing, which is a crucial aspect of natural language processing tasks. The aim is to use a fine-tuned version of SloBERTa to generate paraphrases of sentences by selecting a corpus that includes semantically similar paraphrases while preserving the original meaning. The paper proposes to validate the model using both manual and automatic metrics. The implementation of this approach has the potential to improve the efficiency of natural language processing tasks.

Keywords

Transformer-Based Language Models, SloBERT, paraphrasing, Transformer-Based Language Models

Advisors: Slavko Žitnik

Introduction

Paraphrasing, the process of expressing the same meaning using different words is an important part of natural language processing tasks. Recently, deep learning models, particularly the transformer-based language models like BERT, have achieved remarkable results in several language processing tasks. This paper aims to implement sentence paraphrasing using BERT model, which has demonstrated state-of-the-art performance in range of processing tasks. Specifically we will use fine-tuned version of SloBERTa [1] to generate paraphrases of sentences. Fine-tuning will be performed by one of the selected corpora, which will be preprocessed to include semantically similar paraphrases while preserving the meaning of the original sentence. We will validate our model with manually defined metrics and automatic methods.

Related work and existing solutions

Recently with emergence of ChatGPT service many of other companies and academic groups have speed up development of their own LLMs such as GPT-2 (by OpenAI), GPT-3 (by OpenAI), T5 (by Raffel et al. [2]), RoBERTa (by Liu et al. [3]), LLaMa (by Touvron et al. [4]).

In the article ***The Multilingual Paraphrase Database***[5] the authors expand their English and Spanish paraphrase database to include 21 other languages including Slovene. They extract the paraphrases from bilingual parallel corpora by "pivoting" words and phrases over a shared translation in another language (English). They make the database freely available.

Initial ideas

This section aims to provide initial ideas on how to approach the classroom project.

0.1 Dataset acquisition/generation

Perhaps the most important part of training a good model for paraphrasing is starting with the right dataset. The dataset should include quality pairs of phrases and their paraphrases. Such datasets already exist mainly for the english language such as The Paraphrase Database (PPDB)[6]. We have found an extended version of PPDB that includes other languages including slovene[5]. Upon further investigation we have determined that the phrases included in this extended database are often only synonyms and the "source" phrases are not longer than 6 words. This dataset was produced using the back translation method or by "pivoting" over a shared translation in another language, in this case english. We would like to see if we get better results using a different method such as automatic translation of non-slovene dataset. We will use a translation model provided in the instructions of this assignment [7] to automatically translate the original PPDB to Slovene. Afterwards we will check the quality and similarity of the translated pairs using different metrics. If our method produced better results than the back translation method we would use the automatically translated dataset to train our paraphrasing model.

More recently the QUORA[8] and MSCOCO[9] dataset have been used to train and evaluate paraphrasing models. The QUORA dataset is a dataset of duplicate questions posted on the Quora website. This questions are manually marked by moderators as duplicates so they are confirmed paraphrases. We could use automatic translation to obtain paraphrasing pairs in Slovene. The MSCOCO dataset is not a dataset originally made for paraphrasing or language processing rather it is a dataset of objects and object recognition within images. However the dataset specifically the annotations of the images can still be used to train and test paraphrasing models. To use this model in our assigmnet we would again need to automaticaly translate it and check for the quality of the translated pairs.

0.2 Natural language deep learning model

To implement paraphrasing, we can use a deep learning model to fine-tune RoBERTa model on pairs of sentences. The RoBERTa model is a transformer-based model that is pretrained on a large corpus of text. Fine-tuning the RoBERTa model on pairs of sentences will allow us to train the model to generate paraphrases of the input sentences. The deep learning model will learn to generate paraphrases by minimizing the difference between the predicted paraphrase and the actual paraphrase. This approach has been shown to be effective in generating high-quality paraphrases.

0.3 Result evaluation

The quality of the implemented model has to be somehow evaluated so we can compare it to other already existing models. For this we will use different metrics of evaluation. These are BLEU (*Bilingual Evaluation Understudy)* which was developed to evaluate machine translation systems, but can be also used for paragraph evaluation, METEOR which aims to adress BLEU's weakness of being unable to measure semantic equivalents when applied to low-resource languages and has a better correlation with human judgement at the sentence/segment level than BLEU, GLEU, ROUGE a recall-based evaluation metric originally developed for text summarization, has also been used to evaluate paraphrase generation, BERTscore, WER [10]. And lastly we will use human evaluation as a metric, which will probably be the best in scoring, but most definitely the most tidiest and time consuming.

References

[1] Matej Ulčar and Marko Robnik-Šikonja. Slovenian RoBERTa contextual embeddings model: SloBERTa 1.0, 2020. Slovenian language resource repository CLARIN.SI.

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[5] Juri Ganitkevitch and Chris Callison-Burch. The multilingual paraphrase database. In *LREC*, pages 4276–4283. Citeseer, 2014.

[6] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. Ppdb: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, 2013.

[7] Institute "Jožef Stefan" and Clarin.si. Slovene nmt. https://github.com/clarinsi/Slovene_NMT, 2021. Accessed on March 20, 2023.

[8] Sambit Mishra. First quora dataset. https://www.kaggle.com/datasets/sambit7/first-quora-dataset, 2018. Accessed on March 20, 2023.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[10] Jianing Zhou and Suma Bhat. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086, 2021.