FRI Natural language processing course 2024
Project report

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

Unsupervised Domain adaptation for Sentence Classification

Veronika Matek, Karmen Frank, and Luka Mihelič

Abstract

The abstract goes here. The abstract goes here.

Keywords
Domain Adaptation, TSDAE, GPL, SBERT, Sentence Classification

Advisors: Slavko Žitnik, Aleš Žagar, Boshko Koloski

Introduction

Previous state-of-the-art methods, like SBERT, for deriving sentence embeddings have a key problem of not working for specific topics and domains. We bypass this problem by additionally fine-tuning our non-domain-specific base model using methods like TSDAE and GPL. In this report we fine-tune an unsupervised base model SBERT with the mentioned methods for the classification of Slovenian sentences based on their sentiment. We compare the results between before and after fine-tuning the base model using labeled data and observe the impact of different parameters during the learning of each approach.

Methods

SBERT

SBERT (Sentence Bidirectional Encoder Representations from Transformers) adds siamese and triplet structure networks to the pre-trained transformer network BERT, which produces new state-of-the-art results for NLP tasks such as question answering, sentence classification and sentence-pair regression. SBERT applies a pooling layer to the output of a BERT/RoBERT model, deriving fixed sized sentence embeddings. With the added network structures we can fine-tune the model and update weights so the to output results are sentence embeddings that are semantically meaningful. Semantict aspects embedded in the continuous vector space can be measured with the cosine metric similarity, where similar semantic representations in a high-dimensional vector space are closer to each other. The available training data for a given knowledge domain also defines the SBERT network structure. Therefore we may use the classification, regression or the triplet objective function for different kind of tasks.

CO concatenates the sentence embeddings

Such fine-tuning seeks to produce a fixed sized sentence embedding,

TSDAE

Transformer-based Sequential Denoising Auto-Encoder (TSDAE) is a state-of-the-art unsupervised method used for domain adaptation with an encoder-decoder architecture. A shortcoming of previous sentence embedding techniques like SBERT is the lack of domain knowledge. Fine-tuning a model like this with TSDAE can adapt our model to a specific domain without any labeled data, as this is hard and expensive to acquire [1].

Before training the model, TSDAE corrupts the input sentences, for example by deleting or swapping words, and encodes them to a fixed vector. The goal of the decoder is to reconstruct the vectors to the original input by prediciting what was changed. It is important to note that the decoder has no context as it doesn't have access to other sentence embeddings and thus creates a bottleneck [1]. This architecture can be seen in Figure 1.

For the purpose of classifying Slovenian sentences based on their sentiment we fine-tune the SBERT model with TSDAE. We choose bert-base-uncased (TODO) for our base model to fine-tune. For the loss function we use the Denois-

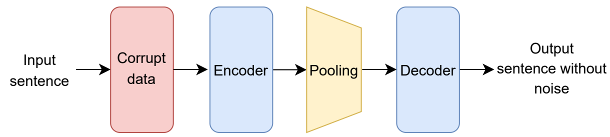

Figure 1. Architecture of TSDAE model.

ingAutoEncoderLoss as our loss function when training. We train the method and compare our results with the corpus.

TSDAE has been shown by Wang, Reimers and Gurevych [1] to outperform other unsupervised approaches and other supervised models, trained with a lot of labeled data. Many previous works were evaluated on Semantic Textual Similarity (STS) which might return good performance but it is unclear how it performs on specific domains.

GPL

The Generative Pseudo Labeling (GPL) is a domain adaptation technique that utilizes unsupervised learning. It allows us to fine-tune a dense retrieval model (for example SBERT [2]) on a desired domain. First step of GPL is preparing (query, sentence)-pairs. This takes three phases: generating suitable queries, negative mining and using cross-encoder to assign a score to each pair [3]. This process is visualised in Figure 2.

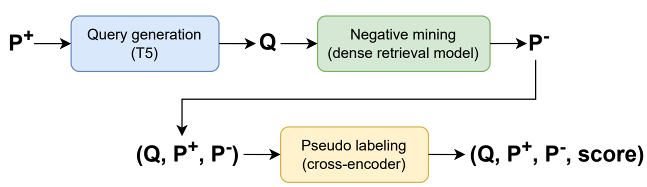

Figure 2. The workflow of GPL's sentence preparation step. Queries Q are generated for each input sentence P^+. The generated queries are then used for negative mining or finding similar sentences P^-. Pseudo labeling step involves a cross-encoder that assigns a score to each (query, sentence)-pair.

Queries are generated using a pretrained T5 encoder-decoder model [4]. Three queries are generated for each input sentence. The next step is negative mining, where 50 of the most similar sentences are retrieved for each of the generated queries, using an existing dense retrieval model. The (query, input sentence)-pairs are denoted as (Q, P^+) and the negative sentence as P^-.

The last step of data preparation involves a cross-encoder that assigns a score to each (query, sentence)-pair. For each (Q, P^+, P^-)-tuple a margin δ is calculated using the next formula:

$$\delta = \mathrm{CE}(Q, P^+) - \mathrm{CE}(Q, P^-), \tag{1}$$

where CE is the score predicted by the cross-encoder. This gives us a dataset $D_{GPL} = \{(Q_i, P_i, P_i^-, \delta_i)\}_i$, which is used for training a dense retrieval model with the MarginMSE loss function. This model thus learns to map queries and sentences into a vector space and is fine-tuned to a given domain.

Data

Kakšne podatke uporabljamo, kako izgledajo, in what way did you prepare the data, delitev na množice (poudarimo, da se vse metode treniranjo z enako učno množico). Pokažemo morda par primerov povedi v tabeli.

Testing approach

Naslov morda še ni ustrezen in se bo prilagodil. Katero metriko uporabimo za primerjavo rezultatov, kako iz sentence embedding pridemo do klasifikacije povedi.

Results

TO DO: Use the results section to present the final results of your work. Present the results in a objective and scientific fashion. Use visualisations to convey your results in a clear and efficient manner. When comparing results between various techniques use appropriate statistical methodology.

Discussion

TO DO: Use the Discussion section to objectively evaluate your work, do not just put praise on everything you did, be critical and exposes flaws and weaknesses of your solution. You can also explain what you would do differently if you would be able to start again and what upgrades could be done on the project in the future.

Acknowledgments

Here you can thank other persons (advisors, colleagues ...) that contributed to the successful completion of your project.

References

[1] Kexin Wang, Nils Reimers, and Iryna Gurevych. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[2] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084, 2019.

[3] Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. GPL: generative pseudo labeling for unsupervised domain adaptation of dense retrieval. *CoRR*, abs/2112.07577, 2021.

[4] 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. *Journal of Machine Learning Research*, 21(140):1–67, 2020.