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. Both of the methods have been proven to significantly outperform previous state-of-the-art models, like Masked Language Model, on domain specific training data, working even better if combined together [1, 2].

Most of these previous successful methods were trained on Semantic Textual Similarity, which does not take into account any domain knowledge. Some examples of these approaches are SBERT and Infersent. One such reason for the lack of domain knowledge is that training a model might require a lot of labeled data, which can be expensive and hard to get. This holds true especially for specific topics. One way of solving this is training the model on the general corpus before fine-tuning it to the required domain [1, 2].

In this report we aim to fine-tune an unsupervised multilingual base model SBERT with two of the mentioned domain adaptation techniques, namely Transformer-based Sequential Denoising Auto-Encoder (TSDAE) and Generative Pseudo Labeling (GPL) on the SentiNews classification dataset. We classify Slovenian sentences based on their sentiment, which can be positive, neutral or negative. We compare the results given by the pretrained base model and by the base model fine-tuned with both mentioned methods as well as a combination of them.

We observe any improvement when faced with domain specific data compared to SBERT, trained on the exact same input sentences. We also distinguish the impact of different parameters during the learning of each approach and try to find the optimums.

Methods

SBERT

SBERT (Sentence Bidirectional Encoder Representations from Transformers) adds siamese and triplet structure networks to the pre-trained transformer network BERT, which produces state-of-the-art results for natural language processing tasks such as question answering, sentence classification and sentence-pair regression. SBERT applies a pooling layer to the output of a BERT/RoBERTa 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 [3]. Semantic 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.

The classification objective function concatenates sentence embeddings u and v with the element-wise difference $|u-v|$ and multiply it with the trainable weight x_1 $W_t \in \mathbb{R}^{3n\times k}$

[3]:

$$o = \text{softmax}(W_t(u, v, |u - v|)) \tag{1}$$

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 size vector. The goal of the decoder is to reconstruct the vectors of 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.

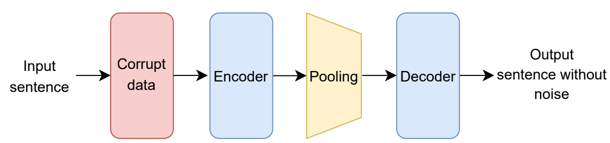

Figure 1. Workflow of TSDAE. The input sentences are first corrupted and then encoded into fixed size vectors. The vectors are pooled and then attempted to be reconstructed with the decoder.

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: change base model accordingly) for our base model. During training we use the DenoisingAutoEncoderLoss as our loss function, which expects pairs of original and corrupted sentences as the input. We train the model where the decoder attempts the reconstruction of the corrupted sentences and compare our results with the corpus [4].

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 [3]) 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 [2]. This process is visualised in Figure 2.

Queries are generated using a pretrained T5 encoder-decoder model [5]. Three queries are generated for each input sentence. The next step is negative mining, where 50

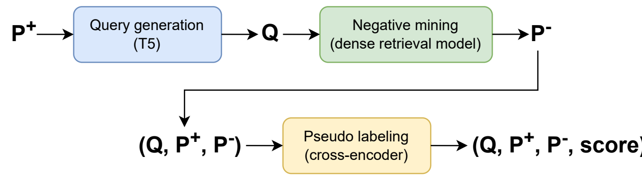

Figure 2. The workflow of GPL's sentence preparation step. Queries Q are generated for each input sentence P^+. The generated queries are then used for negative mining or finding similar sentences P^-. Pseudo labeling step involves a cross-encoder that assigns a score to each (query, sentence)-pair.

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 equation:

$$\delta = \text{CE}(Q, P^+) - \text{CE}(Q, P^-), \tag{2}$$

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.

The MarginMSE loss [6] relies on the scores, or pseudo labels, provided by the cross-encoder. It teaches the dense retrieval model to predict the margin between the score of (Q, P^+)-pair and score of (Q, P^-)-pair. It follows the next equation:

$$\text{MarginMSE} = \frac{1}{N} \sum_{i=0}^{N-1} |\hat{\delta}_i - \delta_i|^2, \tag{3}$$

where N is the batch size, δ_i is defined in equation 3, provided by the cross-encoder, and $\hat{\delta}_i$ is derived by the (student) dense retrieval model, which we are fine-tuning.

Data

We used the SentiNews dataset [7], which contains 169k sentences from 10.4k documents, equiped with sentiment labels, in the Slovenian language. A few examples of dataset's elements are shown in Table 1.

Sentence	Sentiment
Kaže, da se blejskim vilam vendarle obeta lepša prihodnost.	positive
O tem bo Evropska komisija odločala septembra.	neutral
V Sloveniji je ta rast znašala sedem odstotkov.	negative

Table 1. Examples from the SentiNews [7] dataset.

The dataset was split into train, validation and test set (TO DO). For each method we use to fine-tune the base model, the exact same data sets are used. 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 autoencoderfor 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] 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.

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

[4] Sentencetransformers documentation. Accessed: 2024-03-21.

[5] 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.

[6] Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. Improving efficient neural ranking models with cross-architecture knowledge distillation. *CoRR*, abs/2010.02666, 2020.

[7] Sentinews dataset. Accessed: 2024-03-21.