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

Unsupervised Domain adaptation for Sentence Classification

Marko Možina, Peter Kosem, Aljaž Konec

Abstract

TBA

Keywords

Unsupervised Sentence Classification, Generative Pseudo Labeling, Transformer-based Denoising AutoEncoder

Advisors: Boshko Koloski

Introduction

Natural Language Processing (NLP) is most commonly used applications like sentiment analysis, spam detection and topic categorization. Applying general NLP models to specialized domains, where unique terminologies and contexts are needed, can affect model performance. Sentence-transformer models, while effective for generating sentence embeddings, often fall short in these specialized settings without domain-specific tuning.

This study aims to tackle this by enhancing sentence representation in specialized domains through unsupervised domain adaptation techniques [3, 4, 5], specifically Transformer-based Denoising AutoEncoder (TSDAE) [8] and Generative Pseudo Labeling (GPL) [9]. These methods intend to refine the embedding space, making models more sensitive and accurate for specific domains, thereby improving sentence classification outcomes.

The project focuses on Slovenian language and as such will use Slovenian specialized domains SentiNews [2] and siParl [6] for evaluating the adaptation to a specialized domain.

Methods

This section outlines the approach taken to adapt sentence-transformer models for improved sentence classification within specialized domains, leveraging the techniques of Transformer-based Denoising AutoEncoder (TSDAE) and Generative Pseudo Labeling (GPL).

Generative pseudo labeling (GPL)

The ability to effectively process and classify text across diverse domains remains a challenge in natural language processing. Traditional models often fail when applied outside their training domain due to the unique linguistic characteristics of the domain. For instance, the world 'loud' has a positive sentiment if used to describe a speaker and a negative sentiment if used in a review of a hotel. This highlights the need for domain adaptation techniques capable of leveraging unlabeled textual data that is prevalent in specialized fields. Generative Pseudo Labeling (GPL) [9] offers a novel approach to utilize unlabeled data for enhancing model adaptability and performance in specialized, often smaller, domains.

Generative Pseudo Labeling (GPL) is predicated on the innovative use of unlabeled data to improve model functionality in target domains. The GPL methodology unfolds in two stages:

1. **Pseudo Label Generation:** A pre-trained model, proficient in a related but distinct task, assigns provisional labels to unlabeled target domain data. These initial labels, derived from the model's pre-existing knowledge, serve as a foundational step for domain adaptation [7].
2. **Refinement through Generative Modeling:** Subsequently, the model undergoes a self-enhancement phase, refining its capabilities by learning from the data directly. This involves generative models that discern and adapt to the underlying patterns specific to the target domain, thereby aligning the model more closely with the target domain's characteristics.

Our project seeks to leverage GPL for the unsupervised domain adaptation of sentence-transformer models, aiming to bolster sentence classification accuracy within specialized domains. The application process is outlined as follows:

1. **Initial Model Training:** Employing a pre-trained sentence-transformer model, leveraging its extensive knowledge base for a preliminary understanding of the target domain [7].

2. **Pseudo Label Creation:** Generating pseudo labels for the Slovenian classification dataset (e.g., SentiNews) with the pre-trained model, bridging the model's knowledge from general to specific domains.

3. **Model Adaptation via GPL:** A generative model refines the sentence embeddings and classification efficacy of the sentence-transformer, emphasizing the adaptation to capture domain-specific nuances accurately.

4. **Iterative Refinement and Evaluation:** Through continuous refinement and evaluation, the model's performance is iteratively improved, ensuring its alignment with the project's goals.

Transformer-based Denoising AutoEncoder (TSDAE)

The core idea of TSDAE [8] is to introduce noise to input sequences by deleting or swapping tokens (e.g., words). This corrupted input is then fed into the encoder component of the TSDAE, which consists of transformer layers that encode the corrupted input data into a latent space representation of sentence vectors. The decoder network, then aims to reconstruct the original input data from the latent representation. Below is a brief explanation of the sequential process of TSDAE:

1. **Corruption:** The input data is corrupted with noise, introducing variations and disturbances into the data. Adopting only deletion as the input noise and setting the deletion ratio to 0.6 performs best per [8].

2. **Encoding:** The corrupted input data is fed into the encoder, which consists of transformer layers. These layers transform the input data into a latent space representation called sentence vector, capturing essential features while filtering out noise.

3. **Decoding:** The latent representation obtained from the encoder is passed through the decoder, also composed of Transformer layers. The decoder aims to reconstruct the original, clean input data from the latent representation.

4. **Reconstruction:** The classifier token (CSL) embedding is used during reconstruction from token-level to sentence-level representation [1].

5. **Training:** The TSDAE optimizes its parameters by minimizing the reconstruction error between the denoised output generated by the decoder and the original, clean input data. This process occurs iteratively, allowing the model to learn effective denoising strategies.

For fine-tuning the model, we need to set up the training data (which is nothing more than text data, since the model is unsupervised), a pretrained model prepared for producing sentence vectors and a loss function. By leveraging the Transformer architecture, TSDAEs can efficiently capture complex dependencies and patterns in the data, making them effective for denoising tasks across various domains, including natural language processing. Despite its inability to match the performance of supervised methods, TSDAE remains valuable, particularly in scenarios where data is unlabeled or difficult to obtain.

Specialized Domains

As mentioned before, we primarly focus on two specialized domains in Slovenian language: SentiNews [2] and siParl [6]. SentiNews is a Slovenian news dataset containing around 10 427 articles form different news sources. These articles is labeled with sentiment labels using the five-level Lickert scale on three levels of granularity (document, paragraph and sentence).

SiParl is a Slovenian parliamentary dataset containing transcripts of parliamentary sessions. The whole dataset is arounf 11 thousand session, containing 1 million speaches and 200 million words. This dataset is not labeled and we will use it for unsupervised domain adaptation.

References

[1] Unsupervised training for sentence transformers, 2021. (Accessed on 03/21/2024).

[2] BUČAR, J. Manually sentiment annotated slovenian news corpus SentiNews 1.0, 2017. Slovenian language resource repository CLARIN.SI.

[3] GANIN, Y., AND LEMPITSKY, V. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning* (Lille, France, 07–09 Jul 2015), F. Bach and D. Blei, Eds., vol. 37 of *Proceedings of Machine Learning Research*, PMLR, pp. 1180–1189.

[4] KAMATH, U., LIU, J., AND WHITAKER, J. *Transfer Learning: Domain Adaptation*. Springer International Publishing, Cham, 2019, pp. 495–535.

[5] NGUYEN-MEIDINE, L. T., BELAL, A., KIRAN, M., DOLZ, J., BLAIS-MORIN, L.-A., AND GRANGER, E. Knowledge distillation methods for efficient unsupervised adaptation across multiple domains. *Image and Vision Computing 108* (2021), 104096.

[6] PANČUR, A., ERJAVEC, T., MEDEN, K., OJSTERŠEK, M., ŠORN, M., AND BLAJ HRIBAR, N. Slovenian parliamentary corpus (1990-2022) siParl 3.0, 2022. Slovenian language resource repository CLARIN.SI.

[7] REIMERS, N., AND GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[8] WANG, K., REIMERS, N., AND GUREVYCH, I. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (Punta Cana, Dominican Republic,

Nov. 2021), M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds., Association for Computational Linguistics, pp. 671–688.

[9] WANG, K., THAKUR, N., REIMERS, N., AND GUREVYCH, I. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Seattle, United States, July 2022), M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds., Association for Computational Linguistics, pp. 2345–2360.