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) significantly benefits applications like sentiment analysis, spam detection, and topic categorization. The challenge intensifies when applying general NLP models to specialized domains, where unique terminologies and contexts 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 project aims to tackle this issue by enhancing sentence representation in specialized domains through unsupervised domain adaptation techniques, specifically Transformer-based Denoising AutoEncoder (TSDAE) and Generative Pseudo Labeling (GPL). These methods intend to refine the embedding space, making models more sensitive and accurate for specific domains, thereby improving sentence classification outcomes.

By investigating the application of TSDAE and GPL for domain adaptation, this study seeks to bridge the gap between general-purpose sentence embeddings and their effectiveness in domain-specific contexts, offering potential improvements in NLP model performance across specialized fields.

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)

In the vast landscape of digital information, the ability to effectively process and classify text across diverse domains remains a paramount challenge in natural language processing. Traditional models often falter when applied outside their training domain due to the unique linguistic characteristics of new data sets. This gap highlights the urgent need for domain adaptation techniques capable of leveraging the wealth of unlabeled textual data prevalent in specialized fields. Generative Pseudo Labeling (GPL) emerges as a vital solution, offering a novel approach to utilize unlabeled data for enhancing model adaptability and performance in uncharted 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 pivotal 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 [1].
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 [2].

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 [1].

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 [2].

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 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. Decoder network, which also consists of transformer layers, then aims to reconstruct the original, clean input data from the latent representation. Below, we briefly explain 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 [3].

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 [4].

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.

initial Evaluation of TSDAE Models

Objective

This section presents the initial experimental evaluation of various sentence-transformer models, specifically focusing on their performance in sentence classification tasks within the context of unsupervised domain adaptation. The primary objective was to compare the effectiveness of traditional transformer models against their TSDAE-adapted counterparts.

Experimental Setup

The experiments were conducted using three different models: "bert-base-multilingual-uncased", "all-MiniLM-L12-v2", and "hate_speech_slo". Each model was evaluated in its base configuration as well as its TSDAE-modified version. The models were trained on a subset of 10,000 sentences extracted from the SentiNews dataset, aiming to establish preliminary performance metrics.

The dataset was split into training and testing sets with an 80-20 ratio, using a random seed for reproducibility. Model training was followed by sentence embedding generation for both training and testing datasets. Logistic regression was then applied to these embeddings to perform sentence classification, assessing the models based on their F1 scores and log loss metrics.

Results

The performance of each model is summarized as follows:

- **BERT Base Multilingual Uncased**: Achieved an F1 score of 0.606 and a log loss of 0.756. Its TSDAE variant recorded a slightly lower F1 score of 0.604 and a higher log loss of 0.764.

- **Hate Speech SLO**: Mirrored the BERT base model in F1 score (0.606) and log loss (0.756), while its TSDAE adaptation saw a decrease in performance with an F1 score of 0.602 and a log loss of 0.774.

- **All-MiniLM-L12-v2**: The base model had an F1 score of 0.561 and a log loss of 0.820, with the TSDAE adaptation further decreasing in F1 to 0.541 and increasing log loss to 0.826.

The following table illustrates these results in detail:

Model	F1 Score	Log Loss
bert-base-multiling.-uncased	0.606	0.756
tsdae-bert-base-multiling.-uncased	0.604	0.764
hate_speech_slo	0.606	0.756
tsdae-hate_speech_slo	0.602	0.774
all-MiniLM-L12-v2	0.561	0.820
tsdae-all-MiniLM-L12-v2	0.541	0.826

Table 1. Performance metrics of base and TSDAE models.

Analysis

The results indicate a slight performance decrease for TSDAE models compared to their base counterparts across the models tested. This could be attributed to several factors, including the complexity of the domain adaptation task, the characteristics of the dataset, and the inherent challenges in using unsupervised methods for domain-specific sentence classification. Additionally, the TSDAE models were only trained on 10,000 sentences, which may have contributed to their relatively poorer performance due to insufficient training data. Furthermore, the increase in log loss suggests that while TSDAE models may be refining the sentence embeddings, they might also be introducing aspects that do not correlate as effectively with the classification objectives as the original models.

Future Work

Further investigations will focus on optimizing the TSDAE training process, exploring different configurations of noise and corruption techniques, and increasing the dataset size to improve the robustness and accuracy of the models. Additionally, deeper analysis into the types of errors made by the TSDAE models may provide insights into their operational dynamics and potential areas for enhancement.

References

[1] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[2] Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360, Seattle, United States, July 2022. Association for Computational Linguistics.

[3] Kexin Wang, Nils Reimers, and Iryna Gurevych. Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning, 2021.

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