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

Unsupervised Domain adaptation for Sentence Classification

Bernard Kuchler, Jan Šuklje, and Luka Šveigl

Abstract

Advisors: Slavko Žitnik

Introduction

We seek to improve article representation in specialized domain articles by adapting a pretrained model such as SBERT which is not tuned to specific fields. The non tuned model will serve as a baseline to improve upon. We will fine tune the SBERT model on the SentiNews database [1] for articles related to the keywords: Economy, Economics and Bank. We will attempt to accomplish this by using two adaptation techniques: TSDAE (Transformer-based Denoising AutoEncoder) and GPL (generative pseudo labeling).

SBERT, created by Reimers and Gurevych [2] is an improvement upon BERT for the creation of semantically meaningful sentence embeddings. It will serve as a good baseline as it can encode semantics for sentences making it easier to compare them. SBERT presents similar sentences with embeddings which are close together in the vector space while dissimilar sentences are far apart. We will be using a multilingual SBERT model so that we will not require additional training of the model for the Slovene language. A good example for such a model is the Language-agnostic BERT Sentence Embedding (LaBSE) created by Feng et al. [3].

TSDAE, created by Wang et al. [4] is a domain adaptation and pre-training method. TSDAE is used to introduce noise to input sequences by deleting or swapping tokens and turning them into sentence vectors. It then attempts to reconstruct the original input sequence from the vectors. This allows the model to better capture the characteristics of the target domain.

GPL, created by Wang et al.[5] is also a unsupervised domain adaptation method. It consists out of query generation, negative mining and pseudo labeling in a data preparation step and a fine-tuning step.

Methods

Results

Discussion

Acknowledgments

References

[1] Boshko Koloski, Matej Martinc, Ilija Tavchioski, Blaž Škrlj, and Senja Pollak. Slovenian keyword extraction dataset from SentiNews 1.0, 2022. Slovenian language resource repository CLARIN.SI.

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

[3] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding, 2022.

[4] Kexin Wang, Nils Reimers, and Iryna Gurevych. Tsdae: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. *arXiv preprint arXiv:2104.06979*, 2021.

[5] Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. *arXiv preprint arXiv:2112.07577*, 2021.