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

Uvodni zapis - slabosti SBERT-a in podobnih metod, zakaj se uporablja TSDAE in GPL (v povezavi s prilagoditvijo domene). Osredotočili se bomo na "domain adaptation" za slovenski jezik, pri čemer bomo metode testirali nad problemom "sentimental analysis" (nekako povežemo besedilo na problem klasifikacija pri čustvih). Cilj projekta je bila implementacija metod TSDAE, GPL, s katerimi finetunamo SBERT (BERT+pooling), in testiranje metod ter primerjava rezultatov nad problemom klasifikacije stavkov v eno izmed "pos", "neg", "neutral". Opazovali smo tudi izbrane parametre tekom učenja posameznega pristopa.

In the Introduction section you should write about the relevance of your work (what is the purpose of the project, what will we solve) and about related work (what solutions for the problem already exist). Where appropriate, reference scientific work conducted by other researchers.

Methods

Use the Methods section to describe what you did an how you did it – in what way did you prepare the data, what algorithms did you use, how did you test various solutions ... Provide all the required details for a reproduction of your work.

Tu nekoliko opišemo SBERT - predvsem glede sentence embeddingov, ki jih potrebujemo za naprej.

TSDAE

TSDAE - opis metode, pri čemer se predvsem osredotočimo na domain adaptation. Povemo kateri model finetunamo. Dodamo skico arhitekture.

GPL

GPL - opis metode, pri čemer se predvsem osredotočimo na domain adaptation. Povemo kateri model finetunamo. Dodamo skico arhitekture.

The Generative Pseudo Labeling (GPL) [1] is a domain adaptation technique that utilizes unsupervised learning. It allows us to fine-tune a dense retrieval model (for example SBERT) 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.

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