Using Similarity to Select Training Data for Cross-Domain Sentiment Analysis

Chrisanna Cornish
ccor@itu.dk

Danielle Dequin
ddeq@itu.dk

Sabrina Pereira
sabf@itu.dk

Introduction to Problem

Some domains lack substantial resources for training, therefore cross-domain sentiment analysis is still a challenging task. Training a model in Natural Language Processing (NLP) can be a computationally expensive and time intensive task, particularly where data is unlabeled and requires human input to label.

Research Question

In this project, we seek to answer the following research question: can we extend a pre-trained model by selecting additional training data from a target domain based on a similarity criterion? Is there any difference in the amount of additional training data needed when the target domain is similar or dissimilar to the source domain?

Experiment Setup

We will be working with Amazon reviews and product categories will be considered the domains, such as “furniture” or “music”. For this experiment we will be using “music” reviews as our source domain and the target domains will be reviews within the categories “video games” and “arts, crafts, and sewing”. Kullback-Leibler (KL) divergence will be used to measure the similarity between the domains.

In the baseline experiments, we will test how the pre-trained model performs for each target domain, then additional training data will be randomly selected from the target domains to fine tune the model.

The experiments will then be repeated, but instead of selecting random data from the target domains, they will be selected based on dissimilarity to the source domain.

The performance of the model will be measure at each increment, then compared.

What is Novel, Interesting and/or Relevant

To our knowledge, there is no study that has examined specifically the different levels of data required for fine-tuning a pre-trained model in cross-domain sentiment analysis, while comparing the differences when data is randomly versus specifically-selected.

One study uses a Structural Correspondence Learning (SCL) which attempts to learn the co-occurrence between features from the two domains (Blitzer et al., 2007). This method is used to select training data that matches a new domain, but does not have a solution when the source and target domain are dissimilar.

Another study uses Structured Feature Alignment (SFA) which uses some domain-independent words as a bridge between domains (Pan et al., 2010). This method is specifically useful when the target domain is unlabeled, but is not optimal when the domain is labeled and low-resource.

We present an approach of pre-selecting the training data based on similarity to the source data. If this improves performance, it can be used to guide training data selecting in order to fine tune a model in order to improve performance for low-resource domains.

References

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, pages 751–760.