Multilingual Embeddings for Low Density Languages

Advisor: Dragomir Radev

Dennis Zhao

1 Abstract

One of the main challenges in Natural Language Processing (NLP) is understanding low density languages. Low density languages are languages that are low in popularity in resources. One method to deal with low density languages is with zero-shot learning, a method which involves extending supervised learning when not enough labeled examples are available. We seek to use zero-shot learning to create cross-language multilingual embeddings. This will aid in creating embeddings for low density languages, which in turn can help perform tasks such as information retrieval from resources written in low density languages.

2 Background

In recent years, various models for learning cross-lingual representations and creating multilingual embeddings have been proposed [Rud17]. These models generally use four different approaches: Monolingual mapping (initially train on monolingual word embeddings and then learn linear mapping between monolingual representations in different languages), Pseudo-cross-lingual(combine contexts of different languages and then train model to learn cross-lingual relations), Cross-lingual training (train embeddings on parallel corpus), or Joint Optimization (train on parallel and optionally monolingual data and joinly optimize some combination of monolingual and cross-lingual losses).
There are also various types of data that can be used to train models. Depending on the type of data that is available, a different training method may be needed. The data types include Word-aligned data (parallel corpus with word alignments), Sentence-aligned data (parallel corpus without word alignments), Document-aligned data (corpus containing documents in different languages), Lexicon (cross-lingual dictionary with pairs of translations between words), or non-parallel data.

In the case of low-density languages, we often fall into the case of no parallel data or very little parallel data. This serves as our motivation for utilizing zero-shot learning, in order to create quality multilingual embeddings without the need for parallel data. A team for researchers from Carnegie Mellon University (CMU) and the University of Washington have come up with novel dictionary-based methods to create multilingual word embeddings [Amm+16]. Their new method only requires monolingual data and pairwise parallel dictionaries. The model was able to train embeddings in 59 languages; parallel corpora are not required for this model but can be added in if available for the language.

Furthermore, they also introduced a new method of evaluating multilingual embeddings based off of the existing QVEC system and addressed the theoretical shortcomings of multiQVEC. This evaluation tool is now available via web portal [1] which provides a stable metric of multilingual embedding quality for future research in the generation of multilingual embeddings. These methods have also been applied in the Loreilei embeddings [2]. These are a set of multilingual word embeddings for Uzbek, Turkish, Chinese, and Hausa using multiCCA, multiCluster and multiSkip; word2vec was used as the monolingual embedding method.

Another set of multilingual embeddings have been produced by the Polyglot project [APS13]. The Polyglot project is a release of multilingual word embeddings for over 100 languages which were trained on their respective Wikipedias. These embeddings can

be used as off-the-shelf solution to obtain near top performance in fundamental NLP tasks. However, the Polyglot project has not explored using these word embeddings in conjunction with traditional NLP features, which has been shown in other contexts to significantly improve results on NLP tasks.

3 Project

I will be working with Professor Radev as well as other members of the Project MATERIAL team. I will be responsible for initially collecting a list of multilingual NLP resources. I will also implement zero-shot learning to create a set of multilingual embeddings. These embeddings will then be tested with the multiQVEC-CCA, available via web portal. These embeddings will also be used to perform tasks such as information retrieval from resources in low-density languages. However, the specific language that will be used for testing will be released on a later date to ensure that the models are intended to be trained on non-parallel data. The deliverables of this project will include all of the code that I have written, and a report detailing my contributions to the project.

References

