Methods of Mathematical Analysis final project

Polina Stadnikova

February 28, 2018

Abstract

This project implements a basic idea of the inverted indexing approach for creating word embeddings where words are represented as one-dimensional arrays. The objective is to obtain cross-lingual word representations that capture similar or the same concepts in different languages and to investigate potential challenges and limitations. For the evaluation, different techniques are used, including the distance between two vectors. The outcome of the project is a small tool that allows the user to perform some experiments with these embeddings.

1 Method

1.1 Cross-lingual embeddings

Cross-lingual embeddings are projections from different languages into the same semantic space. They can be used in many NLP tasks, such as machine translation, syntactic parsing, POS-tagging, and, therefore, are an object of active research in distributional semantics.

Creating such word embeddings usually requires multilingual parallel corpora. My method follows the idea from [Søgaard et al.] of using Wikipedia as a source.

1.2 Count-based representations

The semantic space of a language can be composed by vectors that correspond to the words in this space. In count-based approaches vectors contain the information about word co-occurrences, represented by raw or weighted counts[Jurafsky & Martin].

1.3 Corpus

1.4 Preprocessing

2 Evaluation

3 User Guide

References

[Søgaard et al.] Anders Søgaard, Zeljko Agic, Hector Martinez Alonso, Barbara Plank, and Bernd Bohnet (2015). Inverted indexing for cross-lingual NLP *In ACL, Vol. 1*, 1713–1722.

[Jurafsky & Martin] Daniel Jurafsky and James H. Martin (2017). Speech and Language Processing. *Third Edition draft*, 275–291.

[Och & Ney, 2004] Och, Franz Josef and Ney, Hermann (2004). The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics 30*, 417–449

[Pasupat & Liang, 2015] Pasupat, Panupong and Liang, Percy (2015). Compositional Semantic Parsing on Semi-Structured Tables. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*,1470–1480.