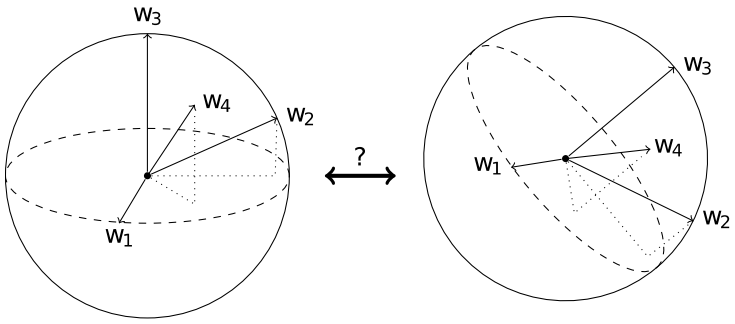
Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Karlsruher Institut für Technologie

Master's Thesis

Shallow Neural Nets - Comparing Diachronic Word Embedding Model Alignments



One major research area is the analysis of the evolution of human language, also referred to as diachronic linguistics. Among other things, words that change their meaning are of particular interest. To this end, word embedding models such as Google's word2vec or Glove are trained for multiple time periods to investigate what words are similar in either period. Yet, the embeddings are not directly comparable unless trained incrementally. This is due to the nondetereminism of learning embeddings. To tackle this problem so called "alignment" approaches have been proposed. By applying various transformations, independently learned models can be "aligned". Thereby, a comparison is re-enabled. However, it is not clear what the different alignments accomplish. Previous work has not analyzed the properties and quality of the alignment approaches. Moreover, it is not even clear whether such an alignment is even feasible.

The focus of this thesis is to compare different alignment approaches for diachronic word embedding models. In particular, the following research questions are of interest:

- How do the embeddings differ when learning multiple times from an identical corpus?
- How can the quality of the alignment be measured?
- Several different approaches exist to learn embeddings. Are the requirements for performing an alignment differing?

This results in the following tasks:

- Theoretical and exploratory analysis on the properties of existing alignment methods
- Design metrics that quantify the quality of the embedding alignment
- Experimental evaluation of different embedding models and alignment approaches on multiple corpora

Throughout this work, you will acquire a deep knowledge of the properties of word embedding models in diachronic analysis. You train highly demanded skills to develop evaluation metrics and conduct controlled experiments to compare state-of-the-art algorithms. Knowledge from a lecture such as "Big Data Analytics" is not a prerequisite. However, elementary statistical knowledge, programming skills and the ability to accomplish conceptual work are desired.

Ansprechpartner

Adrian Englhardt adrian.englhardt@kit.edu +49 721 608-48131 Raum: 341

Am Fasanengarten 5 76131 Karlsruhe Gebäude: 50.34

Master's Thesis