Word Representation for Text Analysis and Search
Document Retrieval, Sentiment Analysis, and Cross Lingual Word Sense Disambiguation

Dipl.-Ing. Navid Rekabsaz BSc.

Abstract

Semantics in language is a fundamental aspect of human cognition and in great extent defines our understanding and knowledge. Word representation methods suggest a computational model to capture semantics by providing vectors as proxies to the meaning of terms, known as word embedding. Recent advancements of the models using neural network approaches open an exciting perspective, and urge further research on understanding and making use of semantic representation models in language and text processing.

In this thesis, we introduce novel methodologies to exploit word representation models in various text analysis tasks. We also provide in-depth analyses of the concept of term relatedness in semantic models. The thesis contributes to basic research in the area of Information Retrieval and word representation interpretability, as well as applied research in Cross-Lingual Word Sense Disambiguation (CL-WSD), and sentiment analysis. We cover several tasks in Information Management such as document retrieval, gender bias detection, CL-WSD for language with scarce resources, and volatility prediction, studied in the news, health, finance, and social science domains.

In the first task—document retrieval—we introduce a novel approach to integrate the information of related terms in traditional retrieval models. The approach generalizes the idea of the translation model to various probabilistic models. In the course of the study, we realize the importance of addressing two relevant topics: how to select the related terms in the representation models, and how to adapt the term similarities to the specific needs of retrieval systems. We approach the former by exploring the space of word vectors, and the latter by combining similarities of representations, created based on different assumptions on the surrounding contexts of terms.
Our evaluations on various retrieval test collections show significant improvements in search performance by using the generalized translation models in comparison to strong, state of the art baselines.

The next topic approaches the interpretability of word embedding by introducing a novel neural-based representation model. The model transfers dense word embedding to sparse vectors where the semantic concepts of the representations are explicitly specified. As a case-study, we use these explicit representations to quantify the degree of the existence of gender bias in the Wikipedia articles. Our analysis shows strong bias in a few specific occupations (e.g. nurse) to female.

The next task regards CL-WSD for low-resource languages/domains (English to Persian in our work). We approach this task using the semantic similarity of the translation terms in their contexts, showing the benefits of exploiting word representation for CL-WSD, specially in the absence of reliable resources.

Finally, we contribute to the state-of-the-art of sentiment analysis, by exploiting the generalized translation models to predict volatility in financial markets. Our approach, when combined with factual market data, outperforms state-of-the-art methods, and shows the advantages of using textual data together with semantic methods for volatility forecasting.

Evaluation Committee

Univ.-Prof. Dr. Allan Hanbury
Ao.Univ.Prof. Dipl.-Ing. Mag. Dr. Margrit Gelautz
Univ.-Prof. Mag. DDr. Stefan Thurner
Prof. Christina Lioma
Prof. Carsten Eickhoff