UNIVERSITÄT BIELEFELD
Faculty of Technology

Bachelor's or Master's Thesis

Word embedding space exploration

Semantic Computing Group
Moritz Blum
mblum@techfak.uni-bielefeld.de

Words that occur in the same contexts tend to have similar meanings. We found that chemical elements that share physical properties (such as being magnetic) occur in similar contexts within natural language texts.
In materials science, a very large database of scientific publications can be accessed. This literature uses a domain-specific vocabulary and is mostly stored in an unstructured way. Having knowledge represented in textual form makes assessing this knowledge difficult. Ideally, relevant knowledge would be made available in the form of a database / knowledge graph.
Word embedding models represent words in a high-dimensional vector space in which similar objects are located close to each other. Moreover, word embeddings reflect analogies like $v_{king} - v_{man} + v_{woman} \approx v_{queen}$.
The goal of this thesis is to use unsupervised learning techniques to find patterns in the vector space and to systematically search for analogies.

Related literature

- De Vine, Lance, Shlomo Geva, and Peter Bruza. "Unsupervised mining of analogical frames by constraint satisfaction," 2018. `https://eprints.qut.edu.au/124091/14/ALTA_2018_devine_et_al.pdf`
- Tshitoyan, V., Dagdelen, J., Weston, L. et al. "Unsupervised word embeddings capture latent knowledge from materials science literature," 2019.
- Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space."
- Dan Jurafsky and James H. Martin: Speech and Language Processing

The Semantic Computing Group researches and develops methods that enable machines to acquire, process, and understand data from natural language and knowledge graphs. Using methods from *natural language processing* and *machine learning*, we are aiming at machines that are capable of knowledge acquisition by reading various kinds of data. In particular, the group focuses on methods for information extraction, knowledge graph completion, semantic parsing, ontology learning, sentiment analysis, entity linking, as well as question answering.

More information is available at:
https://uni-bielefeld.de/fakultaeten/technische-fakultaet/arbeitsgruppen/semantic-computing

Interested? @mail to `mblum@techfak.uni-bielefeld.de`