

Indirectly Knowledge-Informed Word Embeddings

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

Word embeddings are great because they capture semantic relationships and meaning in language, enabling machines to understand and process words in a more meaningful way, facilitating tasks like natural language understanding, sentiment analysis, and machine translation. Background knowledge, such as ontologies, domain-specific information, can provide additional context, that could help to enhance their quality and performance, because the knowledge expressed in texts and the knowledge contained in KGs may complement each other.

For example, the sentence

The Dark Knight is a 2008 superhero film directed by Christopher Nolan.

could be extended into

The Dark Knight (part of the series: The Dark Knight trilogy) is a 2008 superhero film directed by Christopher Nolan (instance of: human).

by inserting background knowledge.

In order to enrich texts with background knowledge about entities, these entities need to be identified in the texts. Thus, named entity recognition and linking need to be carried out. This can be done with Spacy to link to Wikidata. Which knowledge helps most when added to the texts is to be found out in this project. Finally, a model is trained on a plain corpus and a separate model is trained on the enriched corpus. The performance of the model obtained from the enriched corpus is compared against the performance of the model obtained from the plain corpus.

Related literature

- www.spacy.io
- Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
- 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