Labeling Results of Topic Models: Word Sense Disambiguation as Key Method for Automatic Topic Labeling with GermaNet

Jennifer Ecker
IDS Mannheim

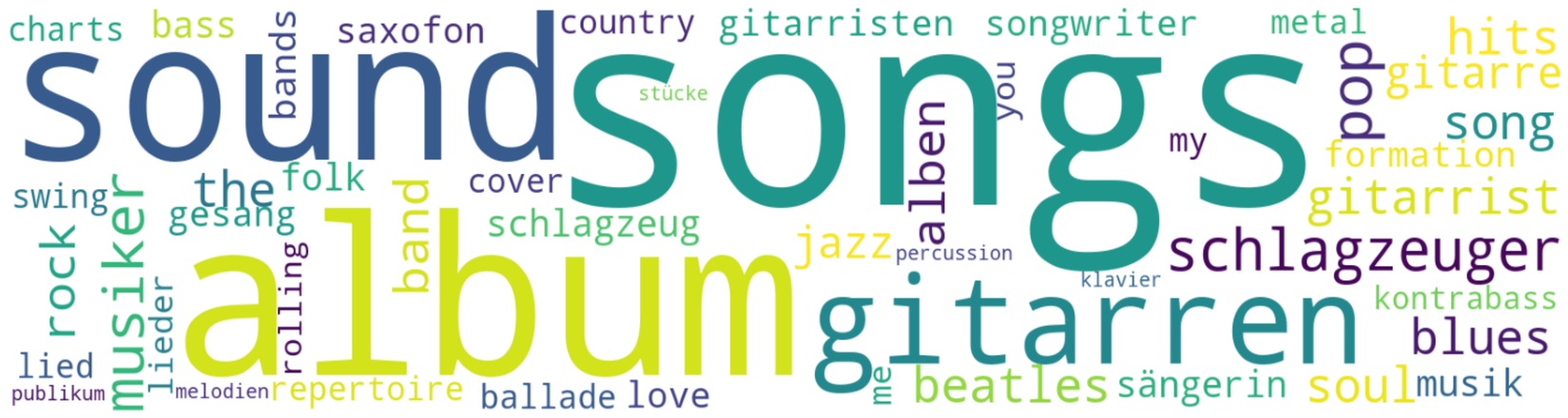
The combination of topic modeling and automatic topic labeling sheds light on understanding large corpora of text. It can be used to add semantic information for existing metadata. In addition, one can use the documents and the corresponding topic labels for topic classification. While there are existing algorithms for topic modeling readily accessible for processing texts, there is a need to postprocess the result to make the topics more interpretable and self-explanatory. The topic words from the topic model are ranked and the first/top word could easily be considered as a label. However, it is imperative to use automatic topic labeling, because the highest scored word is not the word that sums up the topic in the best way. Using the lexical-semantic word net GermaNet, the first step is to disambiguate words that are represented in GermaNet with more than one sense. We show how to find the correct sense in the context of a topic with the method of word sense disambiguation. To enhance accuracy, we present a similarity measure based on vectors of topic words that considers semantic relations of the senses demonstrating superior performance of the investigated cases compared to existing methods.

Data Set

The data set is from the newspaper corpus Mannheimer Morgen Volume 20, a subset of the Mannheim German Reference Corpus (DeReKo). Comprising 44 383 texts, this corpus offers a rich source of linguistic data. Distributed under the QAO-NC license, the data set is provided in I5-formatted XML files.

Resources

Utilizing Top2Vec (Angelov, 2020), a total of 348 distinct topics are extracted from the corpus by employing a *doc2vec* embedding model and the speed parameter *deep-learn*.
The lexical-semantic net GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010) is used to disambiguate polysemous words. For instance, the word *Album* denotes both a music album and a collector's album.

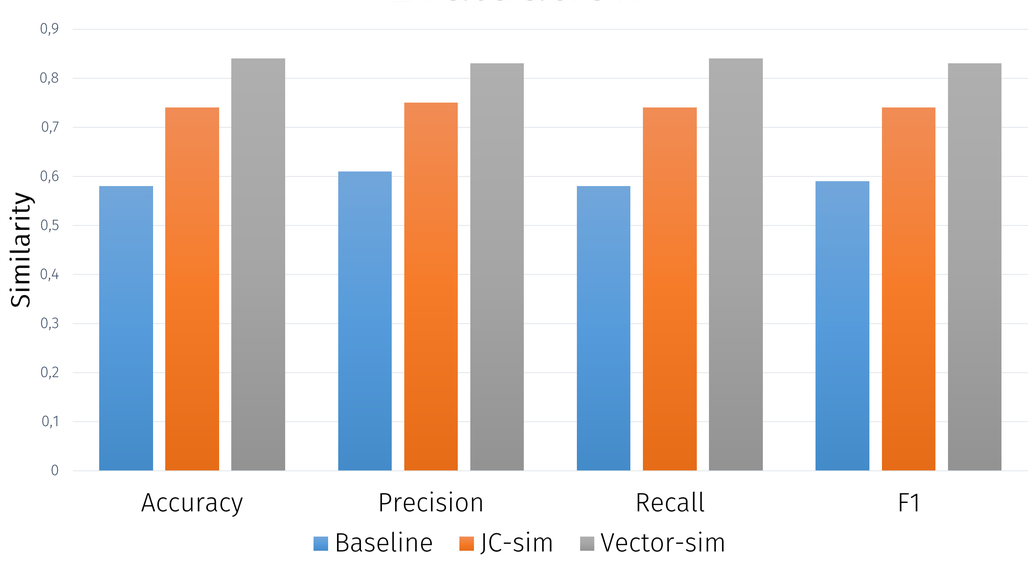


Word cloud of topic 0 from the top 50 words.

Similarity Measures

The Jiang & Conrath similarity (JC-sim) combines path information from a word net with word frequencies from a corpus. We calculate the similarity of the first synset with the first non-ambiguous topic word, then with the second non-ambiguous topic word, and so on. After that, the average of all values for the first synset is calculated. The same procedure is followed with the next synset in order to later compare which average similarity is the highest to choose one synset.
The vector based similarity (Vector-sim) is a combination of using the words and their hypernyms and hyponyms in GermaNet and word embeddings. For this, we use a German word embedding model. And we compare the sum of the vectors of all the topic words that only have one synset (non-ambiguous) with the vector of each sense of an ambiguous word with their direct hypernyms and direct hyponyms. The comparison is made with the cosine similarity between the mean vectors of two sets. We choose the sense with the highest cosine similarity.

Evaluation



Results

We demonstrate the efficiency of utilizing topic words for disambiguating polysemous words. Disambiguation can be achieved without relying on complete sentences, particularly in scenarios where contextual sentences (glosses) are absent in a word net. We show the effectiveness of leveraging the structural relationships within GermaNet, such as hypernyms and hyponyms, to enhance the disambiguation of polysemous words. Using vectors as similarity measure is superior to corpus-based information content calculations like the JC-similarity, leading to higher accuracy and F1-scores.

Contact
Jennifer Ecker
Digital Linguistics
Leibniz Institute for the German Language
R5, 6-13
68161 Mannheim
Germany

https://www.ids-mannheim.de/digspra/
ecker@ids-mannheim.de

References

Angelov, Dimo. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.
Hamp, Birgit and Feldweg, Helmut. 1997. Germanet-a lexical-semantic net for german. In *Automatic information extraction and building of lexical semantic resources for NLP applications*.
Henrich, Verena and Hinrichs, Erhard W. 2010. Gernedit-the GermaNet editing tool. In ACL (System Demonstrations), pages 19–24.