Linguistics Society of India

Paper title:
Automatic Synset Ranking for Indian Languages using Word Embeddings

Abstract

Synset ranking is an important step towards achieving word sense disambiguation (WSD). Often, the acid test for any WSD approach is to beat the first sense heuristic, as listed by WordNet. However, many approaches, even the aforementioned skyline is dependent on the availability and the quality of annotated datasets and resources such as Wordnet.Thus, such approaches do not perform well in case of languages with incomplete wordnets. In this paper, we propose an unsupervised approach to synset ranking which does not rely on any annotated data. It utilizes word embeddings, the quality of which is dependent on the raw corpus on which they are trained. We argue that for a resource scarce language, raw corpora are better than their annotated corpora and wordnets in terms of completeness, coverage, etc. Thus our approach is bound to work well even for such languages. We justify our claim by evaluating this approach on multiple languages viz., English, Hindi, Marathi, Bengali, Kannada, Punjabi etc. Our results show that this approach is promising.

References:

- Atreya, A., Kakde, Y., Bhattacharyya, P., and Ramakrishnan, G. (2013). Structure cognizant pseudo relevance feedback. In the 6th International Joint Conference on Natural Language Processing (IJC-NLP), Nagoya, Japan.
- Balamurali, A. R., Joshi, A., and Bhattacharyya, P.(2012). Crosslingual sentiment analysis for indian languages using linked wordnets. In International Conference on Computational Linguistics (COLING), Mumbai, India. Citeseer.
- Bhattacharyya, P. (2010). Indowordnet. In Language Resources and Evaluation Conference (LREC)
- Bhingardive, S., Shaikh, S., and Bhattacharyya, P. (2013). Neighbors help: Bilingual unsupervised wsd using context. In Association for Computational Linguistics (ACL), Sofia, Bulgaria
- Bhingardive, S., Singh, D., V, R., Redkar, H. H., and Bhattacharyya, P. (2015). Unsupervised most frequent sense detection using word embeddings. In NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015, pages 1238–1243