Extracting lexical causatives from discourse

Guanghao You, Moritz Daum, Sabine Stoll
University of Zurich
guanghao.you@uzh.ch, daum@psychologie.uzh.ch, sabine.stoll@uzh.ch

Keywords: language processing, first language acquisition, causatives, machine learning, discourse genres

Causation is one of the main features of human cognition and language. An important step in language acquisition is to understand causation and its linguistic expressions. A prerequisite for this is the extraction of causatives from the input. Languages vary in how they express causatives but the main three types are lexical, periphrastic and morphological causatives. While periphrastic and morphological causative constructions can usually be easily traced by detecting periphrastic verbs (e.g. “make” in English) and affixes (e.g. -(s)ase in Japanese), lexical causatives have no explicit marker and are therefore much more difficult to generalize. Essentially, verbs can imply different levels of causality, which might form a continuum rather than exhibiting strict cutting points for lexical causatives. Here we propose a computational method to simulate the extraction and generalization of lexical causatives based on distributional learning. To test whether child-directed speech exhibits a different distribution of features which might facilitate this generalization process we test our method in three different corpora representing three genres: written language, spoken adult-to-adult language and child-directed speech.

We employ the word embeddings algorithm (Mikolov et al., 2013) with an adjustable window to generate high-dimensional vector representations for verbs. Besides inferring from raw utterances in the corpora, the models capitalize on syntactic information at different levels to achieve a more comprehensive inference. We apply the models to corpora of three different discourse genres: the written corpus in the British National Corpus (BNC), dialogues in the BNC spoken corpus and child-directed speech in the Manchester Corpus (Theakston et al., 2001). Our main finding is that syntax plays a different role for causative inference in different genres. In the written corpus, results show that models built solely on raw utterances generally perform below the baseline, while adding syntactic information did improve the performance. In particular, post-verbal parts of speech (POS) help to largely enhance the differentiation in most cases. Moreover, the order of feeding syntactic information to the models in the training phase drastically influences the performance, with feeding syntax after raw utterances worsening the results across all window sizes. Adult-to-adult dialogues show no clear patterns concerning the inference strategy of lexical causatives but often display improved results when syntax is involved. While syntactic information is favored in the two genres taken from the BNC, child-directed speech consistently renders a relatively high performance purely based on raw utterances. We suggest that verb meaning in child-directed speech is most successfully extracted by relying on the semantics of neighboring words whereas meaning in written language is best extracted by relying on structural information of the neighboring words. We discuss the impact of these results for first language acquisition research and the potential generalization of these models in other languages.

References

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language, 28*, 127-152.