Donnerstag
09.03.2017
11:15 – 11:45
B3 2, 0.03

Uniform Information Density models for language production: A comparative study of Hindi and English

Ayush Jain, Vishal Singh, Sumeet Agarwal, Rajakrishnan Rajkumar
(*Indian Institute of Technology Delhi*)

{ee1120439, ee1120493, sumeet, raja}@iitd.ac.in

AG1

In this paper, the extent to which language production is governed by the Uniform Information Density (UID) hypothesis (Jaeger 2010) is analysed for Hindi and English. The hypothesis states that a speaker tries distributing the information across a sentence in the most uniform fashion possible so as to communicate efficiently, analogously to communication in a noisy channel (Shannon 1948). We examine the effect of word-order change on the information distribution of a sentence and its effect on production choice.

Several quantitative UID models were defined, based on the variation of lexical *N*-gram scores within a sentence. As alternatives, the UID measures were also estimated using syntactic data (part-of-speech tags) instead of lexical data, and only at chunk boundaries rather than at all words. In order to assess the validity of the UID hypothesis for language production, the information orthogonal to that captured by a basic *N*-gram score was estimated by performing a binary classification (reference corpus sentences vs. variants) and comparing the classification accuracy in the presence and absence of UID scores.

While using lexical UID measures, segregation of corpus and non-corpus sentences was poor for both English and Hindi. However for syntactic UID measures, the segregation was much better for English, though for Hindi it was still poor. On addition of normalized UID measures to a baseline feature set consisting of *N*-gram scores, the classification accuracy for the Hindi sentences increased slightly, suggesting that perhaps the notion of UID might in some cases be informative about production choices even for Hindi. The difference in the results obtained for Hindi and English is possibly because of greater flexibility of word order in Hindi.

References: • T. Florian Jaeger. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62, 2010. • C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948. • C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(4):623–656, Oct 1948.