Large-Scale Paraphrasing for Natural Language Understanding

Chris Callison-Burch[2,3] Benjamin Van Durme[1,2] Matt Post[2]
Juri Ganitkevitch[1] Jonathan Weese[1] Ellie Pavlick[3]

[1]Center for Language and Speech Processing
Johns Hopkins University

[2]Human Language Technology Center of Excellence
Johns Hopkins University

[3]Computer and Information Science Department
University of Pennsylvania

We present an overview of our recent and ongoing paraphrasing work. Our work encompasses large-scale extraction of syntactically annotated paraphrase pairs, which we learn from bilingual parallel corpora using the intuition of "pivoting" over foreign-language expressions: we assume two English expressions that translate to the same foreign expression to be meaning-equivalent.

Further, we utilize monolingual text corpora to collect distributional signatures for English phrases. This allows us to annotate the bilingually extracted paraphrases with an additional signal based on contextual similarity. As a result of this effort we release the paraphrase database PPDB, a collection of millions of automatically extracted and ranked paraphrases.

Additionally, we present a domain adaptation scheme for paraphrasing that relies on extracting paraphrases from only the parts of the general-domain data that are most similar to a sample of target domain data.

We also outline NattyLo, a project that will classify the extracted paraphrase collection into more fine-grained relation categories, like forward- and backward-entailment, aiming for better performance in tasks like recognizing textual entailment. NattyLo also includes the development of an entailment recognition approach based on parsing with synchronous context-free grammars.