

Jenny Loomis
jloomis@stanford.edu
DHCS 2009 Proposal

Metaforager: A Pattern-Learning System for Large-Scale Metaphor Extraction

This paper explores the possibilities of Dual Iterative Pattern Relation Expansion (Brin, 1998; extended by Agichtein, 2000) in extracting metaphors from literature. Previous metaphor extraction systems generally used supervised learning techniques (e.g. Pasanek, 2008) and/or relied on hand-built databases of word relations (such as Mason, 2004, and other WordNet-based efforts). Employing information retrieval algorithms more commonly used for structured relation extraction, and influenced by the scalability constraints of commercial data mining, Metaforager is a highly scalable, minimally-supervised learning system for extracting metaphors from large plain-text corpora.

In the current prototype, a 20-node Hadoop cluster iterates over a corpus of approximately 5,000 books. The system is seeded with a handful of pairs of words which rarely co-occur outside of a metaphor (e.g. "love, fire"). Each iteration consists of a sequence of two MapReduce processes: in the first, the corpus is searched for co-occurrences of the word pairs. Resulting phrases are used to construct patterns based on the words surrounding the pair. These patterns then serve as the input to a second map-reduce task which searches the corpus for a new set of word pairs. The cycle then begins again, using the new pairs in place of the seed pairs to generate another set of patterns.

Since the negative effects of false positives compound with each iteration, this approach depends heavily on effective culling of undesirable terms at every step. This is done by calculating a confidence level for each term; the higher the confidence, the more likely it is that that term is part of a metaphor. Only terms with confidence above a preset threshold are retained. The confidence of a given search term is computed based on the overlap of the results it generates with the set of terms already accepted. Search results which weren't found in the accepted set during the query's confidence calculation comprise the next set of search terms. Tests currently under way measure relative performance (in terms of both precision and recall) achieved by variations in pattern construction, confidence calculation and confidence threshold.

References

- Agichtein, Eugene and Luis Gravano. (2000). Snowball: Extracting relations from large plain-text collections. *Proceedings of the 5th ACM International Conference on Digital Libraries*.
- Brin, Sergey. (1998). Extracting patterns and relations from the World-Wide Web. *Proceedings of the 1998 International Workshop on the Web and Databases (WebDB'98)*.
- Fellbaum, Christiane, editor. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Mason, Zachary J. (2004). CorMet: a computational, corpus-based conventional metaphor extraction system. *Computational Linguistics* 30(1):23-44.
- Pasanek, Bradley M. and D. Sculley. (2008) Mining Millions of Metaphors. *Literary and Linguistic Computing*, 23(3):345-360.