T7: On-Demand Distributional Semantic Distance and Paraphrasing

Yuval Marton

ABSTRACT

Semantic distance measures aim to answer questions such as: How close in meaning are words A and B? Fore example: "couch" and "sofa"? (very); "wave" and "ripple"? (so-so); "wave" and "bank"? (far). Distributional measures do that by modeling which words occur next to A and next to B in large corpora of text, and then comparing these models of A and B (based on the "Distributional Hypothesis"). Paraphrase generation is the task of finding B (or a set of B's) given A. Semantic distance measures can be used for both paraphrase detection and generation, in assessing this closeness between A and B. Both semantic measures and paraphrasing methods are extensible to other textual units such as phrases, sentences, or documents.

Paraphrase detection and generation have been gaining traction in various NLP subfields, including:

- Statistical machine translation (e.g., phrase table expansion)
- MT evaluation (e.g., TERp or Meteor)
- Search, information retrieval and information extraction (e.g., query expansion)
- Question answering and Watson-like applications (e.g., passage or document clustering)
- Event extraction / event discovery / machine reading (e.g, fitting to existing frames)
- Ontology expansion (e.g., WordNet)
- Language modeling (e.g., semantic LM)
- Textual entailment
- (Multi-)document summarization and natural language generation
- Sentiment analysis and opinion / social network mining (e.g., expansion of positive and negative classes)
- Computational cognitive modeling

This tutorial concentrates on paraphrasing words and short word sequences, a.k.a. "phrases" -- and doing so overcoming previous working memory and representation limitations. We focus on distributional paraphrasing (Pasca and Dienes 2005; Marton et

al., 2009; Marton, to appear 2012). We will also cover pivot paraphrasing (Bannard and Callison-Burch, 2005).

We will discuss several weaknesses of distributional paraphrasing, and where the state-of-the-art is. The most notable weakness of distributional paraphrasing is its tendency to rank high antonymous (e.g., big-small) and ontological sibling (e.g., cow-sheep) paraphrase candidates. What qualitative improvement can we hope to achieve with growing size of monolingual texts? What else can be done to ameliorate this problem? (Mohammad et al., EMNLP 2008; Hovy, 2010; Marton et al., WMT 2011).

Another potential weakness is the difficulty in detecting and generating longer-than-word (phrasal) paraphrases, because pre-calculating a collocation matrix for phrases becomes prohibitive in the matrix size with longer phrases, even with sparse representation. Unless all phrases are known in advance, this becomes a problem for real-world applications.

We will present an alternative to pre-calculation: on-demand paraphrasing, as described in Marton (to appear 2012). There, searching the monolingual text resource is done on-demand with a suffix array or prefix tree with suffix links (Manber and Myers, 1993; Gusfield, 1997; Lopez, 2007). This enables constructing large vector representation, since there is no longer a need to compute a whole matrix. Searching for paraphrase candidates can be done in a reasonable amount of time and memory, for phrases and paraphrases of an arbitrary maximal length. The resulting technique enables using richer -- and hence, potentially more accurate -- representations (including higher-dimension tensors). It opens up a great potential for further gains in research and product systems alike, from SMT to search and IR, event discovery, and many other NLP areas.