Comparing two corpus-based methods for extracting paraphrases to dictionary-based method

ABSTRACT

Paraphrase extraction plays an increasingly important role in language-related research and applications in areas such as information retrieval, question answering and automatic machine evaluation. Most of the existing methods extract paraphrases from different types of corpora by using syntactic-based approaches. Since a syntactic-based approach relies on the similarity of context to identify and capture paraphrases, other than paraphrases, other terms which tend to appear in a similar context such as loosely related terms and functionally similar yet unrelated terms tend to be extracted. Besides, different types of corpora suffer from different kinds of problems such as limited availability and domain biased. This paper presents a solely semantic-based paraphrase extraction model. This model collects paraphrases from multiple lexical resources and validates those paraphrases semantically in three ways; by computing domain similarity, definition similarity and word similarity. This model is benchmarked with two outstanding syntactic-based approaches. The experimental results from a manual evaluation show that the proposed model outperforms the benchmarks. The results indicate that a semantic-based approach should be applied in paraphrase extraction instead of a syntactic-based approach. The results further suggest that a hybrid of these two approaches should be applied if one targets strictly precise paraphrases.

Keyword: Paraphrase extraction; Semantics; Lexical resource; Word similarity; Sentence similarity; Domain similarity