Abs-210

Gintarė Grigonytė (Vytautas Magnus University), Algirdas Avižienis (Vytautas Magnus University, University of California Los Angeles), and Rūta Marcinkevičienė (Vytautas Magnus University)

Automatic detection of semantically related lexical items in domain corpora

We present our experimental findings of a new unsupervised approach for the detection of semantically related lexical items by aligning paraphrases in monolingual domain corpora. Two methodologies are applied: automatic extraction and alignment of paraphrases (Cordeiro et. al 2007), and automatic synonymy detection in paraphrased corpora (Grigonyte et. al 2010). Both methodologies were applied to two domain corpora: ReSIST corpus in English on computer security and dependability (Avizienis et. al 2008) and ŠIMTAI corpus in Lithuanian on education and research policy. In the English corpus semantically related lexical items were identified with a 67.27% precision.

Semantically related lexical items can be identified on the basis of their context (Evert 2005). However, the drawback of distributional similarity approach is its dependency on the specific context where items have to be frequently used. To overcome this bottleneck we propose to align paraphrases from domain corpora and to detect lexical items that could be mutually replaceable within the aligned context, called paraphrase casts patterns (Grigonytė et. al 2010). The difference from the pattern-based approach is that instead of looking for pattern-alike lexical items, the local general environment, or discourse, is used as some sort of pattern. This methodology is language-independent, it also does not depend on linguistic processing, or manual definition of patterns as well as training sets, therefore it guarantees a higher precision when compared to distributional similarity-based approaches.

The process of the extraction of semantically similar word pairs or phrase pairs is as follows: domain texts – paraphrased sentences – aligned paraphrased corpus – paraphrase cast examples —– examples of semantically related lexical items. The processing pipeline starts with unstructured domain texts. All the sentences are compared and if similarities are detected, the alignments are established. The method highlights identical or almost identical sentences, it also takes into account pairs of sentences that have a high degree of lexical reordering.

The proposed approach uses Sumo-metric described by Cordeiro et al. (2007) that outperforms simple N-gram overlap, edit-distance and BLEU metric to calculate the semantic similarity of the sentences aligned. When the paraphrases are detected and similar sentences aligned, specific segments, i.e. paraphrase casts, are explored. With the help of paraphrase casts patterns, single word or multi-word lexical items are extracted. Some of extracted pairs appear to be synonymous (advisory vs. expert institution), others have different semantic relationship i.e. that of hyper- and hyponymy (overall education vs. professional training), equanimy (qualification of a psychologist vs. qualification of a teacher) or antonymy (senior vs. junior researcher). A considerable amount of the identified pairs have one word in common without any other specific semantic relation (a roadmap vs. a final version of the strategy).

References

Avižienis, A., Čulo, O., Grigonytė, G., Marcinkevičienė, R.: Building a Thesaurus and Ontology of the Concepts of Dependability and Security. In proc. of 37th IEEE/IFIP International Conference on Dependable Systems and Networks, 2007, p. 420-421.

Cordeiro, J.; Dias, G. & Cleuziou, G. 2007. Biology Based Alignments of Paraphrases for Sentence Compression. In Proceedings of the Workshop on Textual Entailment and Paraphrasing (ACL-PASCAL

/ ACL2007).

Evert, S. 2005. The Statistics of Word Co-occurrences: Word Pairs and Collocations, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Grigonytė, G, Cordeiro, J.P, Moraliyski, R. Dias, G. and Brazdil, P. Paraphrase Alignment for Synonym Evidence Discovery, Proceedings of the 23rd Int. Conf. on Computational Linguistics, COLING 2010, p.403-411.