Abs-132

Ismaïl El Maarouf (Université Bretagne Sud, UEB)

Verb-noun collocations at the crossroads of discourse surface patterns

Collocation has been widely studied in Corpus Linguistics (Firth, 1957; Palmer, 1968). It refers to significant lexical patterns throwing light on word use and meaning. Traditional automatic collocation analysis (Sinclair, 1966, 1991; Sinclair et al., 2004) uses word units, word spans ("windows"), and statistic metrics to retrieve frequent or significant collocations for a given word. Such an approach heavily depends on statistic measures (Mutual Information, Z-score; Church & Hanks, 1989; Clear, 1993) to select and order relevant collocates.

Since then, various techniques have been proposed to classify collocations automatically. One contribution is the use of syntactic data to filter collocates (Kilgarriff et al., 2004). Such an approach classifies collocations according to syntactic relations such as Subject or Object between a noun and a verb. Based on a hand-crafted grammar, the system returns the total number of words satisfying a grammar rule in a given corpus. Thus, the issue of window parameters does not hold explicitly in this approach, but the selection of collocates is dependent on the parser's success i.e. grammar coverage.

In both approaches, text distance between nodes and collocates remains a problem. Regarding the traditional approach, a significant collocate may well lie outside the fixed span while an "irrelevant" collocate might be counted inside the span, though it does not hold any particular relation with the node. The syntactic approach is not foolproof either, one reason being that grammars are not designed to deal with discourse phenomena. Grammars generally make use of finite-state automata to define rules on a limited context (Evert & Kermes, 2003) but there is little work on how those rules interact with discourse structure (see however Say & Akman, 1997, Bayraktar et al., 1998). For instance, phrases separated by commas are not related to each other, with the outcome that a verb separated from its subject by interpolated material is not retrieved. In this perspective, we may expect precision and recall drops for collocation extraction (Kilgarriff et al., 2010).

Our approach (drawing on Jones, 1996) consists in identifying and organizing sentence blocks according to a set of punctuation and discourse markers before syntactic analysis, in order to deal with discourse variation phenomena, such as interpolated clauses and phrases. In a second step, the parser analyses those pseudo-blocks where material irrelevant for the task of relation detection has been discarded. This method has the benefit of filtering beforehand irrelevant collocates, limiting parser's errors and selecting new candidates for collocation.

The paper presents a quantitative analysis of the types of blocks found by our system in a large Press corpus and investigates the benefits of this method with respect to the Subject relation collocates.

References

Bayraktar M., Say B. & V. Akman, 1998, "An Analysis of English Punctuation: The Special Case of Coma". In *IJCL*, 3(1): 33-58.

Church K.W. & P. Hanks, 1989, "Word Association Norms, Mutual Information, And Lexicography". In *Computational Linguistics*, 16(1) : 22-29.

Clear, J., 1993, "From Firth Principles — Computational Tools for the Study of Collocation". In Baker M., Francis G. & E. Tognini-Bonelli (eds.), 1993, *Text and Technology*.

Evert S. & H. Kermes, 2003, "Annotation, storage, and retrieval of mildly recursive structures". In

Proceedings of SProLaC 2003.

Firth J.R., 1957, *Papers in Linguistics 1934-1951*. London: Oxford University Press.

Jones B., 1996, *What's The Point? A (Computational) Theory of Punctuation*. Phd Thesis, University of Edimburgh.

Kilgarriff A., Rychly P., Smrz P. & D. Tugwell, 2004, "The Sketch Engine". In *Proceedings of Euralex 2004*.

Kilgarriff A., Kovář V., Krek S., Srdanović I. & C. Tiberius, 2010, "A Quantitative Evaluation of Word Sketches". In *Proceedings of Euralex 2010*.

Palmer F. R. (eds), 1968, *Selected Papers of J. R. Firth, 1952-1959*. London: Indiana.

Say B. & V. Akman, 1997, "Current Approaches to Punctuation in Computational Linguistics". In *Computers and the Humanities*, 30(6). pp 457-469.

Sinclair J McH, 1966, "Beginning the study of Lexis". In Bazell C. E. , Catford J. C. , Halliday M. A. K., R. H. Robins (eds), 1966, *In Memory of J. R. Firth*. pp. 410-430.

Sinclair J. McH, 1991, *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair J. McH, Jones S., R. Daley, 2004 (1970), *English Collocation Studies – The OSTI Report*. R. Krishnamurthy (eds). London: Continuum.