Wikipedia-based methods to identify noun compounds in running texts

István Nagy T.

In natural language processing, multiword expressions (MWEs) have been received special interest. Noun compounds (NCs) form a subtype of multiword expressions: they form one unit the parts of which are meaningful units on their own, and the unit usually has some extra meaning component compared with the meanings of the original parts [2]. The semantic relation between the parts of the noun compound may vary: it may express a "made of" relation (*apple juice*), a "location" relation (*neck pain*), a "made for" relation (*hand cream*) just to name a few. Thus, noun compounds encode some important meaning relations that can be fruitfully applied by e.g. information extraction systems. However, such applications require that noun compounds should be previously known to the system.

Noun compounds are very frequent in language use (in the Wiki50 corpus [2] 67.3% of the sentences contain a noun compound on average). Furthermore, they are productive: new noun compounds can enter the language all the time hence they cannot be exhaustively listed and appropriate methods should be implemented for their identification. In this paper, we introduce several methods to automatically identify noun compounds on the basis of Wikipedia, like dictionary labeling, rule based methods and machine learning based approaches and show how the expansion of Wikipedia helps the performance of different NC-detecting methods.

References

[1] Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expres- sions: A Pain in the Neck for NLP. In: Proceedings of the 3rd International Confer- ence on Intelligent Text Processing and Computational Linguistics (CICLing-2002, Mexico City, Mexico (2002) 1-15.

[2] Vincze, V., Nagy T., I., Berend, G.: Multiword expressions and named entities in the Wiki50 corpus. In: Proceedings of RANLP 2011, Hissar, Bulgaria (2011).