Extracting Word Collocations from Textual Corpora

Martin PLANK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova, 842 16 Bratislava, Slovakia
plank09@student.fiit.stuba.sk

Natural language is the main way of communication between people. They use it for asking and answering questions, expressing opinions, beliefs, as well as talking about events etc. And they communicate in natural language on the Web, too. However, the simplicity of creating the Web content is not only the advantage of the Web, but also its disadvantage. It is expressed in natural language, which means that it is usually unorganized and unstructured. This makes processing of the Web content expressed in the natural language difficult.

Difficulties in natural language processing are often connected with ambiguity of the language. Some words have specific meaning, when they are used together in one sentence. This raises the problem of collocation extraction. Detection of collocations is important for various tasks in natural language processing (word sense disambiguation, machine translation, keyword extraction etc.). Many statistical methods, as well as other natural language attributes (e.g., part of speech) are used to resolve this task.

Pecina [3] argues that natural language cannot be simply reduced to lexicon and syntax. Individual words can be combined in various ways. This fact is common for most natural languages. The term collocation has several definitions. Choueka [1] defines a collocational expression as "a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components".

During the last 30 years, several association measures were proposed for automatic collocation extraction. The most of the methods are based on verification of typical properties of collocations [3]. It is possible to mathematically describe these properties and determine the degree of association between the components of a collocation. These formulas are called association measures. They compute association score between all collocation candidates in a corpus. The score indicates the likelihood that a candidate is a collocation. These measures can be used for candidate ranking or for classification (if there is a threshold).

* Supervisor: Marián Šimko, Institute of Informatics and Software Engineering

Spring 2013 PeWe Workshop, April 5, 2013, pp. 103–104.

Advantage of these methods is that they can be combined together and final association score can be computed using several measures. Pecina [1] compares 84 statistical measures for collocation extraction. The best results are achieved by the pointwise mutual information. He proposes also method, which finds linear combination of selected methods that improves the performance significantly.

Other approaches employ methods based on the linguistic properties of collocations (e.g., [4]). Manning and Schütze [2] describe the three properties:

- Non-(or limited) compositionality. The meaning of a collocation is not a straightforward composition of the meanings of its parts. For example, the meaning of ‘red tape’ is completely different from the meaning of its components.
- Non-(or limited) substitutability. The parts of a collocation cannot be substituted by semantically similar words. Thus, ‘gut’ in ‘to spill gut’ cannot be substituted by ‘intestine’.
- Non-(or limited) modifiability. Many collocations cannot be supplemented by additional lexical material. For example, the noun in ‘to kick the bucket’ cannot be modified as ‘to kick the {holey/plastic/water} bucket'.

Wermter and Hahn [4] present method based on the non-(or limited) modifiability. The method is built on assumption, that context of a collocation is particularly characteristic. They experiment with this method and compare it to the basic statistical methods. Proposed method significantly outperforms these statistical methods.

In our work we focus on extracting collocations in the Slovak language. We analyze several methods for collocation extraction. Our goal is to adapt or improve existing methods and explore collocation properties in the Slovak language. The important choice is whether to focus on statistical methods measuring co-occurrence between word n-grams or linguistic methods. The statistical methods might be simpler, but one the other hand, they are not based on linguistic collocational properties. Wermter and Hahn [4] show, that exploring these properties is reasonable approach, too. In addition, these methods are able to outperform statistical methods.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0971/11.

References

[1] Choueka, Y.: Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in Large Textual Databases. In: *Proceedings of the RIAO*, CID, 1988, pp. 609-624.
[2] Manning, C.D., Schütze, H.: *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999.
[3] Pecina, P.: An extensive empirical study of collocation extraction methods. In: *Proceedings of the ACL Student Research Workshop*. ACLstudent '05, Association for Computational Linguistics, 2005, pp. 13-18.
[4] Wermter, J., Hahn, U.: Collocation extraction based on modifiability statistics. In: *Proceedings of the 20th international conference on Computational Linguistics.* COLING '04, Association for Computational Linguistics, 2004.