Collocation Extraction on the Web

Martin PLANK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
plank.martin@gmail.com

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.

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).

Other approaches employ methods based on the linguistic properties of collocations. Manning and Schütze [2] describe these characteristic properties:

- Non-(or limited) compositionality. The meaning of a collocation is not a straightforward composition of the meanings of its parts.

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

Spring 2014 PeWe Workshop, March 21, 2014, pp. 33–34.

- Non-(or limited) substitutability. The parts of a collocation cannot be substituted by semantically similar words (synonyms).
- Non-(or limited) modifiability. Many collocations cannot be supplemented by additional lexical material.

In our work, we propose a novel method based on the limited modifiability of collocations. The modifiability of a combination of words can be computed according to the frequencies of n-grams. The method can be explained on a simple example of collocation *to pull my leg* (to tell me something untrue):

1. The frequencies of collocation candidate and its components are computed. Also, the headword (important in the next steps) is identified (*leg*). Headword is one of the collocation components, which has a high semantic significance.
2. Frequent bigrams, which contain the headword, are identified (*right leg, long leg,* etc.). We call the certain number of the most frequent bigrams *headword supplements*. Their frequencies are computed, too.
3. The candidate is modified by the headword supplements. The result is a list of candidate modifications: *to pull my right leg, to pull my long leg*, etc.
4. The frequencies of headword supplements and candidate modifications are compared and the modifiability of a collocation candidate is computed.

The process of judging the collocation candidate is based on the following hypothesis: If the frequencies of candidate modifications are significantly lower than the frequencies of the original candidate, its components and the headword supplements, the candidate is probably a collocation. In other words, if the candidate can be supplied by other lexical information, it has high modifiability. Otherwise, if it cannot be supplied, it has low modifiability and is probably a collocation.

In the experiments, we compared our method with the state-of-the-art association measure *pointwise mutual information*. We compared precision and recall of these methods. Both methods achieved comparable results.

Extended version was published in Proc. of the 10th Student Research Conference in Informatics and Information Technologies (IIT.SRC 2014), STU Bratislava, 161-166.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

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.