Method for Automatic Collocation Extraction from Ukrainian Corpora

Maria Kuzmina, Svitlana Petrasova

National Technical University "Kharkiv Polytechnic Institute",
Pushkinska str., 79/2, Kharkiv, Ukraine

marika958034@gmail.com, svetapetrasova@gmail.com

The article deals with the methods for automatic collocation extraction from Ukrainian corpora. The task of collocation extraction is considered in terms of a corpus-oriented approach [1], based on statistical measures. The term "collocation" is defined as a non-random combination of two words that go together regularly.

Nowadays, the interest in collocation studies relates to the developing of new search engines and machine translation systems, technologies for text recognition, and is due to high frequency of word combinations in text corpora.

The important task of modern corpus linguistics is the extraction of relevant linguistic information, in particular through the use of statistical methods. Additionally, corpus studies allow verifying linguistic theories and hypotheses, as well as identifying and interpreting new language facts [2].

Therefore, the problem of word compatibility or collocation extraction from text data is one of the up-to-date challenges in corpus linguistics.

Linguistic information is extracted and analyzed from the corpus using special programs, namely, corpus managers. The well-known general-purpose corpus managers are SARA, XAIRA (BNC), CQP, which are designed to search data and obtain statistical information from a corpus.

To extract collocations, various association measures can be used. For instance, MI, PMI, t-score, Dice measure, Log-Likelihood are applied to calculate the degree of closeness between components of word combinations in a text corpus. The main drawbacks of statistical methods are noise extraction and ignoring of syntactic correlations between words in long distances.

Analyzing statistical methods for collocation extraction [3], the MI measure has been chosen as a tool for automatic collocation extraction in the Ukrainian-language corpus. MI (Mutual information) measure is used to determine the occurrence of two words by comparing the frequency of their co-occurrence with the product of frequencies of their independent occurrence in the text [4]. One of the advantages of MI is that it allows highlighting key terms that characterize the subject area.

For extracting collocations a corpus of technical instructions has been developed. The corpus meets such requirements as: representativeness, balance, selection, machine readability and standard. The corpus includes three categories (subcorpora) of instructions for existing mobile devices of certain companies. Each text file (instruction) consists of 16 000-17 000 words. The total volume of the developed

COLINS'2018, Volume II: Workshop. Lviv, Ukraine, June 25-27, 2018, ISSN 2523-4013
http://colins.in.ua, online

corpus is 197 624 words. The proposed algorithm for extraction of collocations from the developed text corpus is as follows:

1. Calculating the total number of word forms in the corpus.
2. Defining absolute frequencies of all the words.
3. Defining the frequency of bigrams.
4. Calculating the closeness value for word pairs using the MI measure. To normalize the values, the MI measure is modified by using the MI3 metric (raise the value to the third power).
5. Calculating dispersion and identifying the threshold for selecting the bigrams: +0.5 from the minimum value.

The designed implementation allocates two-word collocations of several types: terminological and general-language word combinations, proper names, phrases that characterize the topic of the text, as well as some free collocations (fig.1).

```
('датчик', 'наближення'):36193.37
('наближення', 'придбайте'):57365.14
('сторонніх', 'гарнітур'):84038.41
('гарнітур', 'аксесуарів'):36193.37
('аксесуарів', 'металевими'):36193.37
('металевими', 'брелоками'):36193.37
('вплинути', 'прийом'):93558.51
('задня', 'кришка'):36193.37
```

Fig. 1. Implementation of MI3 for Collocation Extraction

The extraction of collocations as statistically significant units allows automating the processing of natural-language information, as well as obtaining data about the mechanisms of phrase formation for their further analysis.

All the mentioned methods do not allow receiving the entire range of collocations from analyzed texts. Thus, there is a need to improve the technologies for collocation extraction from natural-language texts, particularly Ukrainian-language texts. In the future research we intend to broaden the scope of the study on collocation extraction using both statistical and syntactical approaches.

References

1. Khohlova M.V. Experimental verification of methods for collocation extraction // Corpus approaches. – Helsinki. – 2008. – P. 343-357.
2. Bobkova T. Classification of Collocation: The Main Approaches and Criteria // Respectus Philologicus. – No. 29(34). – 2016. – P. 87–98. [in Russian]
3. Petrasova S.V., Khairova N.F. A logical and linguistic model for identification of collocation similarity // Bulletin of NTU “KhPI”. Series: System analysis, control and information technology. – Kharkiv, NTU “KhPI”, 2015. – No. 58 (1167). – P. 14–17.
4. Evert S., Krenn B. Methods for the qualitative evaluation of lexical association measures // Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. – Toulouse, France, 2001. –P. 188–195.