Identify of the Substantive, Attribute, and Verb Collocations in Russian Text

Julia Lytvynenko[0000-0001-6087-8155]

National Technical University «Kharkiv Polytechnic Institute»,
2, Kyrpychova str.,61002, Kharkiv, Ukraine

julialytvynenko12@gmail.com

Abstract. This article describes methods and existing libraries for POS-tagging and collocations extraction, using NLP technologies, processing natural language text in the Python programming language. In addition, it describes one of the possible methods for the selection of collocations for a given pattern.

Keywords: POS-tagging, Python, collocation, corpus linguistics, collocations extraction, morphological marking, computer linguistics, intellectual technologies

There is plenty of different purposes of using certain collocations in text or corpus of text. Collocations play an important role in lexicography (the whole collocation dictionaries are created), they are used in ontology compilation, clusterization, language learning, and some other NLP applications [1].

In our case, this task is part of the development of a program for the identification of closely related text fragments; which may be useful for information retrieval. We have consistently distinguished substantive, attribute, and verb combinations.

Collocation means the co-occurrence of two words in some defined relationship. [2]. We look at several such relationships, including direct adjacency and first word to the left or right having a certain part-of-speech. Currently, the term “collocation” is widely used in corpus linguistics, in which the concept of collocation is rethought or simplified compared with traditional linguistics. This approach is can be called statistical. The frequency of joint occurrence is given a priority, so collocations in corpus linguistics can be identified as statistically stable phrases [1].

To date, scientists have created and considered many different methods for isolating collocations. Among them are statistical methods [4] (association measures, t-score measures etc.), as well as methods based on linguistic models. This idea is laid out and implemented in the well-known system Sketch Engine [3]. It gives typical for a given keyword phrases due to, on the one hand, - syntax, that imposes a restriction on the compatibility of words in a given language, and, on the other hand, probabilistic regularities associated with semantics and linguistic patterns.

In our case, we have created a corpus consisting of approximately 20,000 words, where articles on the subject of information technology are collected. This corpus underwent morphological marking (POS-tagging), on the basis of which we could

select the required collocations.

POS-tagging is an automatic morphological markup, which results in each word being tagged. Their values and attributes are determined by the morphological information of each word. For the implementation of morphological marking, the following methods are used: non-verbal morphology, vocabulary morphology based on the base vocabulary, vocabulary morphology on the basis of wordform dictionary, morphemic analysis, Mark chain method, and the N-gram method. Among the existing libraries for POS-tagging in Python there are systems: TreeTager, Pymorphy2, nltk [5].

For morphological processing, we used pymorphy2 and nltk. At the beginning of text processing, we need to normalize the text data and divide the text into tokens. The library nltk is best suited for this. After that, we can start our POS-tagging by using pymorphy2, because it is the best for processing texts of Ukrainian and Russian origin.

When we managed to get the marked text, we need to write the necessary collocations in three files according to three patterns:

1. substantive (NOUN + NOUN(in genitive case)) collocations;
2. attribute (NOUN + ADJECTIVE, which have the same case, number and gender);
3. verb (VERB or INFINITIVE (transitive) with NOUN(in accusative case)

To do this, we create three methods. Each will find the corresponding template. In these methods we again use nltk and pymorphy2. We consider each sentence separately, we look for the first match with the pattern, and if the first is found, we look for a pair for it. If all the conditions are met, we get a list of collocations in the resulting file.

As a result of our text processing, we got 3 lists of collocations, corresponding the template. From the corpus of 20000 words, we have got 668 attributive collocations, 2754 substantive and 452 verb collocations. For example, we got such attributive collocations: краткий обзор, учебно-методический комплекс, учебного заведения etc." We also got substantive collocations: "форм подготовки, обзор исследований, направления подготовки, система подготовки, использованию технологий etc." Among verb collocations were: "имеет специфику, исследовать технологию, предполагает организацию etc."

At the same time, we get a small percentage of collocations that were mistakenly chosen from the text, like substantive collocations: "дисциплине средства, Европе организации, обучения 60-е, годы века etc." The error arises from the fact that the selected collocation does conform to the pattern, but the words themselves, although met in the same sentence, do not have the indicated type of connection between themselves. This problem can be solved by upgrading the algorithm and adding methods, which can help to determine the syntactic links between the words of the text.

The next step we plan is identifying synonymous collocations or the we can put the results in the collocation dictionaries.

References

1. Khokhlova, M., Zakharov, V.: Study of effectiveness of statistical measures for collocation extraction on russian texts. Computer linguistics and intellectual technologies, 9 (16), pp. 137-143 (2010) In Russian.
2. Yarowsky, D.: One sense per collocation. In: 93 Proceedings of the workshop on Human Language Technology, pp. 266-271 (1993)
3. Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D.: The Sketch Engine. In: Proceedings of the Eleventh EURALEX International Congress. Lorient, pp.105–116 (2004)
4. Fano, R.: Transmission of Information, Cambridge (MA)(1961)
5. Jurafsky, D.: Speech and Language Processing, Stanford University, University of Colorado at Boulder, 558p. (2018)