Using a Distributional Semantic Model for Collocation Identification

Anna Mosinyan, Svitlana Petrasova[0000-0001-6011-135X]

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

annymosinyan@gmail.com, svetapetrasova@gmail.com

Abstract. This paper proposes the approach to automatic collocation identification using both the distributional semantic model and POS-tagging. The authors suggest calculating PMI to obtain a sequence of collocations from the designed corpus of research abstracts. Then POS-tagging is applied to classify collocations extracted from the text corpus.

Keywords: distributional semantics, POS-tagging, collocation, text corpus, research abstracts.

Our study analyses models of distributional semantics for solving the problem of automatic identification of collocations in a text corpus. Distributional semantics is an area of linguistics that deals with the computation of the degree of semantic closeness between linguistic units based on their distribution in large arrays of linguistic data (text corpora).

The most widely known distributional semantic models are the following:

1) vector space models;
2) latent semantic analysis;
3) topic models;
4) dependency vectors, etc. [1].

Thus, distributional semantics includes the analysis of co-occurrence of tokens in large text corpora. Distributional analysis allows identifying word combinatory, i.e. collocations. Broadly speaking, a collocation is a combination of two or more words that tend to co-occur. In a narrow sense, a collocation is a typical combination of words, the simultaneous occurrence of which is based on the regular nature of mutual expectation.

Depending on semantics and functions of their use, collocations are classified as follows: traditional (common), expressive, occasional, ethno-cultural, and terminological.

The syntax deals with three main types of collocations: verbal (the main word is expressed by a verb or adverb), substantive (a noun is the main word), and adjective

COLINS'2020, Volume II: Workshop. Lviv, Ukraine, April 23-24, 2020, ISSN 2523-4013
http://colins.in.ua, online

(an adjective is the main word).vCurrently, there are several ways to calculate the combinatory of collocates [2]. Mathematical tools for establishing a syntagmatic connection between words in the text are association measures. We consider one of the measures such as PMI (pointwise mutual information) that has become a mainstream research paradigm in corpus linguistics. This measure of association quantifies the discrepancy between the probability of likelihood of collocates in joint distribution and their independent individual distributions.

Generally, services for collocation extraction apply the statistical apparatus. These include: a bigram search engine, corpus search service using CQP corpus manager, Sketch Engine and others.vIn our study, identifying collocations we combine the use of POS-tagging and the distributional semantic model of PMI.

The algorithm includes the following steps:

1. Create a corpus of abstracts from scientific articles on the ScienceDirect platform.
2. Use POS-tagging to identify parts of speech in the text corpus.
3. Calculate PMI for identifying collocates.
4. Classify collocations based on POS-tagging (carried out in Step 2 of the algorithm).

An example of collocation classification is shown in Table 1.

Table 4. Example of extracted collocations

Type of collocations	Headword	Collocate
Verbal	Verb	Noun
	make	decision
	develop	argument
Substantive (Objective)	Noun	Noun
	name	brand
	account	executive
Substantive (Attributive)	Noun	Adjective
	illness	mental
	care	medical

The further work will be directed at the implementation of the designed algorithm to extract lists of key collocations from scientific texts of a particular domain.

References

1. Lenci, A.: Distributional Models of Word Meaning. Annual Review of Linguistics, 2018. 4:1, pp. 151-171. Available at: https://www.annualreviews.org/doi/abs/10.1146/annurev-linguistics-030514-125254, last accessed 2020/04/07.
2. Petrasova, S., Khairova, N., Lewoniewski, W., Mamyrbayev, O., Mukhsina, K.: Similar Text Fragments Extraction for Identifying Common Wikipedia Communities. Data. MDPI AG, Basel, Switzerland, 2018. 3(4), 66. Available at: https://www.mdpi.com/2306-5729/3/4/66, last accessed 2020/04/07.