DOI: 10.15439/2018F186
ISSN 2300-5963 ACSIS, Vol. 15

Automatic Extraction of Synonymous Collocation Pairs from a Text Corpus

Nina Khairova*, Svitlana Petrasova*, Włodzimierz Lewoniewski†, Orken Mamyrbayev‡ and Kuralai Mukhsina§
*National Technical University "Kharkiv Polytechnic Institute",
Kyrpychova str., 61002, Kharkiv, Ukraine
Email: khairova@kpi.kharkov.ua, svetapetrasova@gmail.com
†Poznań University of Economics and Business,
Al. Niepodległści 10, 61-875 Poznań
Email: wlodzimierz.lewoniewski@ue.poznan.pl
‡Institute of Information and Computational Technologies,
125, Pushkin str., 050010, Almaty, Republic of Kazakhstan
Email:morkenj@mail.ru
§Al-Farabi Kazakh National University, Kazakhstan,
71 al-Farabi Ave., Almaty, Republic of Kazakhstan,
Email: kuka_aimail.ru

***Abstract*—Automatic extraction of synonymous collocation pairs from text corpora is a challenging task of NLP. In order to search collocations of similar meaning in English texts, we use logical-algebraic equations. These equations combine grammatical and semantic characteristics of words of substantive, attributive and verbal collocations types. With Stanford POS tagger and Stanford Universal Dependencies parser, we identify the grammatical characteristics of words. We exploit WordNet synsets to pick synonymous words of collocations. The potential synonymous word combinations found are checked for compliance with grammatical and semantic characteristics of the proposed logical-linguistic equations. Our dataset includes more than half a million Wikipedia articles from a few portals. The experiment shows that the more frequent synonymous collocations occur in texts, the more related topics of the texts might be. The precision of synonymous collocations search in our experiment has achieved the results close to other studies like ours.**

I. Introduction

OVER the last few years, there has been an upsurge of interest in the research which focuses on ways to the retrieval and identification of semantic similarity for textual elements of various levels (words, collocations, and short text fragments). One of the main reasons for this is the expansion of the boundaries of the use of semantically similar texts fragments in various natural language processing applications. Nowadays, words similarity can be processed in Information retrieval systems, Question answering systems, Natural language generation systems, Plagiarism detection systems, Automatic essay grading systems and some others. The second reason for the growth of interest in the identification of a semantic similar element in texts is that on social media billions of small text messages are made public every day, each of which is comprised of approximately thirty words. Whereas very popular traditional algorithms, such as, for example Tf-idf used to compare texts, often fail in very short texts [1]. For this reason, sometimes semantic algorithms and techniques are more needed than statistical ones.

Now there exist enough studies concerning the problems related to the search of words with similar meaning. We could divide all the existing approaches into two groups. The first group of studies is based on the relations of the concepts in a thesaurus. The second group of methods for computing word similarity is based on the appliance of distributional models of meaning.

Measuring the semantic similarity between sentences or collocations is a more challenging task than searching words with similar meaning. Since the task of deciding whether two sentences or two collocations express a similar or identical meaning requires a deep understanding of the meaning of the text fragment. Increasingly, this task is being integrated into the common challenges of the paraphrases [2].

II. Related work

The most explored level of text similarity for the different languages is the level of words. There are a lot of different approaches and methods of computing words similarity. Some of them use thesaurus relations of hyponyms or hypernyms to compute word similarity; the others use distributional similarity of words in a corpus.

However, automatic synonymous collocation pairs extraction from corpora is the more challenging task of NLP. As the task involves two simultaneous operations. The first operation is the collocations extraction from a corpus and the second one is the acquisition of their synonymous pairs.

Wu and Zhou [3] suggested a method that firstly gets candidates of synonymous collocation pairs based on a monolingual corpus and then selects the appropriate pairs of the candidates using their translations in a second language. Pasca and Dienes

This work was supported by the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan

[4] offered to utilize the alignment of two sentences fragments in order to retrieve small phrases with the same meaning. Barzilay and McKeown [5] like [3] built upon the methodology developed in Machine Translation. They presented an unsupervised learning algorithm for identification of similar phrases from a corpus of multiple English translations of the same source text.

Increasingly, the task of the synonymous collocation pairs extraction is being integrated into the common challenge of the paraphrases, which is interpreted as the search of the various textual realizations of the same meaning. Typically, n-gram models [2], annotated corpora and bilingual parallel corpora [6], [7] are used for paraphrases in such studies. Han et al. [8] and Kenter [9] are some of the most recent studies that concern determining the semantic similarity between short fragments of texts. Han et al. [8] combined lexical similarity features, Latent Semantic Analysis (LSA) similarity using WordNet knowledge, alignment algorithm and support vector regression model and n-gram models in order to establish the semantic text similarity. Kenter performed semantic matching between words in two short texts and used the matched terms to create a saliency-weighted semantic network [9].

In our study, we propose using logical equations in order to search collocations of similar meaning in English texts. These equations are based on conjunctions of morphological and semantic characteristics of the words that constitute the collocations. In order to correctly identify the grammatical characteristics, we exploit Stanford POS tagger and Stanford Universal Dependencies (UD) parser[1]. Additionally, in order to pick synonymous words which constitute the collocation we use WordNet synsets[2].

In order to evaluate our approach, we use Wikipedia articles from a few projects. Traditionally, articles of Wikipedia cover various subjects. However, depending on a topic and language versions, Wikipedia community has different numbers of experienced authors or experts [10]. Such groups of users often work together within some subject area of Wikipedia project. Articles related to the projects can have a specific writing style and quality standards, which are defined by the user community of these projects. Therefore, we can expect there is a lot of synonyms and synonymous collocations in texts related to similar topic.

III. Logical-linguistic model

According to previous studies [11], [12], the proposed logical and linguistic model formalizes semantically similar elements of a text by means of grammatical and semantic characteristics of words in collocations.

The semantic-grammatical characteristics determine the role of words in substantive, attributive and verbal collocations. Defining a set of grammatical and semantic characteristics of collocation words, we use two subject variables a^i and c^i. In substantive, attributive and verbal collocations, a set of possible semantic and grammatical characteristics for the main collocation word is defined by the predicate *P(x)*, for the dependent collocation word it is defined by the predicate *P(y)*.

The two-place predicate *P(x,y)* describes a binary relation which is a subset of the Cartesian product of $P(x) \wedge P(y)$ and so determines a correlation of semantic and grammatical information of collocation words *x* and *y*:

$$\begin{aligned}
&P(x,y) = y^{NObjAtt}x^{NSubAg} \vee (x^{NObjOfAg} \vee \\
&x^{NObjOfAtt} \vee x^{NObjOfPac} \vee x^{NObjOfAdr} \vee \\
&x^{NObjOfIns} \vee x^{NObjOfM})y^{NObjAtt} \vee \\
&x^{VTr}y^{NObjPac} \vee y^{AAtt}(x^{NSubAg} \vee \\
&x^{NObjAtt} \vee x^{NObjPac} \vee x^{NObjAdr} \vee \\
&x^{NObjIns} \vee x^{NObjM}) \vee x^{NSubAg}y^{APr}
\end{aligned} \tag{1}$$

Using the algebra of finite predicates, we define the value of the predicate of semantic equivalence for three main types of collocations:

$$\begin{aligned}
&\gamma(x_1,y_1,x_2,y_2) = (x_1^{NSubOfAg} \vee x_1^{NSubAg}) \wedge \\
&y_1^{NObjAtt}(x_2^{NSubOfAg} \vee x_2^{NSubAg})y_2^{NObjAtt} \vee \\
&x_1^{VTr}y_1^{NObjPac}x_2^{VTr}y_2^{NObjPac} \vee x_1^{NSubAg} \wedge \\
&(y_1^{AAtt} \vee y_1^{APr})x_2^{NSubAg}(y_2^{AAtt} \vee y_2^{APr})
\end{aligned} \tag{2}$$

IV. The stages of our methodology

In order to show the correctness of our synonymous collocations extraction model we have used methodology that comprises a few steps. Fig. 1 shows the structural scheme of the methodology, which includes POS-tagging phase, Stanford UD parser and exploitation of the lexical database WordNet.

In the first phase, we employ POS-tagging and UD parser to define the grammatical and semantic characteristics of words in sentences.

The main reason to use UD parser is that its treebanks is centrally organized around notions of subject, object, clausal complement, noun determiner, noun modifier, etc. [13]. Therefore the syntactic relations which connect words of a sentence to each other can express some semantic content.

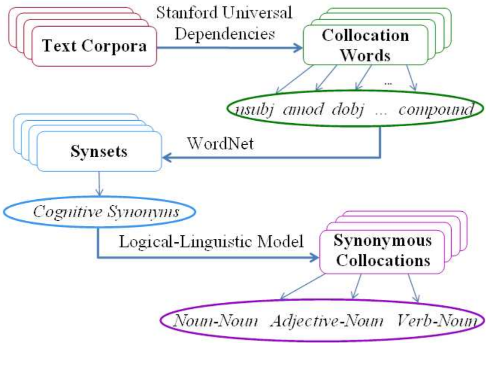

Fig. 1. The structural scheme of our experiment methodology

[1] http://universaldependencies.org/

[2] http://www.nltk.org/_modules/nltk/stem/wordnet.html

We took six types of syntactic relations tags *(compound, nmod, nmod:possobj, obj (dobj), amod* and *nsubj*) from a fixed inventory of UD grammatical relations to denote directed relations between two nouns, a verb and a noun, a noun and an adjective.

Grammatical and semantic characteristics realized through the syntactic relations tags correspond to the variables value in the two-place predicate of equation (2).

Next phase, we use WordNet in order to obtain synonyms of words connected with these types of the syntactical relations. For each collocation (substantive, attributive and verbal), synonyms are searched in WordNet synsets.

If a synonymous word is found, conformity of grammatical and semantic characteristics of a collocation and a potential synonymous word combination is being checked using the proposed logical-linguistic model.

Table I shows the examples of identified synonymous collocations in *Art* and *Biography* Wikipedia portals.

V. Source data and experimental results

Our dataset includes more than half a million articles (502 274) from Wikipedia belonged four thematic projects related to two portals. We focused on projects and portals of Wikipedia because they constitute a huge corpus of texts, which are combined by a common subject, and, at the same time, these texts are written by various authors and, consequently, may contain a lot of different synonyms.

In our studies, we choose two of the biggest portals: *Art* and *Biography*. Each portal can consist of different Wikiprojects. For our experiments, we choose four Wikiprojects (two projects from each selected Wikipedia portals)[3].

In order to estimate our synonymous collocations extraction model, we focus on three approaches. In the first approach, we identify synonymous collocations in any Wikiproject. In the second approach, we identify synonymous collocations in two different projects of the same portal. In the third of our experiments, we identify synonymous collocations in two different projects of two different portals. Our hypothesis is that two projects of the same portal may have the higher number of synonymous collocations than two projects that belong to different portals. The hypothesis is based on an idea that synonyms are occurring more often in related topics texts.

Based on the Corpus Linguistics approaches [14], in order to have the opportunity to compare the synonyms occurrence frequency in the Wikiprojects of different sizes, we normalized the frequencies per ten thousand words. Additionally, we devoted attention to synonymous collocations distribution by three types.

Tables II - IV show relative frequencies of synonymous collocations that occur in four different Wikiproject, two different projects of the same portal and two different projects of two different portals, respectively.

TABLE II
Relative frequencies of synonymous collocations that occur in the Wikiproject

Wikiproject	**The relative frequency of synonymous collocations**		
	Substantive	Attributive	Verbal
Album	214.4	144.8	12.9
Film	277.5	281.7	10.9
Politics and government	200.9	175.4	3.5
Science and academia	280.2	210.4	210.4 4.6

TABLE III
Relative frequencies of synonymous collocations that occur in two different projects of two different portals

Wiki-projects / portal	**The relative frequency of synonymous collocations**		
	Substantive	Attributive	Verbal
Album – Film / Art	199.3	166.3	5.9
Politics and government - Science and academia /Biography	200.8	162.5	3.9

The results of Tables II - III show that the number of synonymous collocations in articles belonging to one Wikiproject is more than the number of synonymous collocations in articles belonging to two different Wikiprojects.

The results of Tables III - IV show that the number of synonymous collocations in articles belonging to the one portal is more than the number of synonymous collocations in articles belonging to two different portals. The articles of one project are closer to one subject than the articles of two different projects.

However, Wikiprojects can also have similar fields of knowledge. Due to it, articles from different projects of the same portal might have enough synonymous collocations.

TABLE I
The Examples of synonymous collocations extracted from Art and Biography portals

Collocations	**Syntactic relation tags**	**Synonymous collocations**	**Syntactic relation tags**	**Collocation types**
history of land	nmod:of	nation's story	nmod:poss	Substantive
soul power	compound	ability of person	nmod:of	Substantive
spectacular progression	amod	outstanding advance	amod	Attributive
restoration is incompetent	nsubj:cop	restitution is incapable	nsubj:cop	Attributive
qualify place	dobj	modify position	dobj	Verbal
preserve fire	dobj	maintain flame	dobj	Verbal

[3] List of the articles of each Wikiproject was extracted in April 2018 https://tools.wmflabs.org/enwp10/cgi-bin/list2.fcgi

TABLE IV
RELATIVE FREQUENCIES OF SYNONYMOUS COLLOCATIONS THAT OCCUR IN TWO DIFFERENT PROJECTS OF THE SAME PORTAL

Wiki-projects (portal)	**The relative frequency of synonymous collocations**		
	Substantive	Attributive	Verbal
Album (Art)–Politics and government (Biography)	128.7	117.2	2.8
Album (Art)–Science and academia (Biography)	172.5	144.4	3.8

VI. EVALUATION RESULTS AND CONCLUSIONS

In our experiments, we use precision to assess the validity of our approach. The main reason why we could not evaluate recall of experiment results that we did not have a training corpus with correctly identified synonymous collocations.

In order to obtain the number of correctly found semantically similar collocations, we use an expert opinion. About 1000 synonymous pairs of collocations were randomly extracted from each list of three types of collocations and presented for judgment. The purpose of the evaluation was to obtain judgments on how synonymous collocations found in the texts were similar in meaning. The experts were asked to compare the similarity of meaning of the collocation pairs on the scale of from 0 to 2. The experts needed to assess the pair of collocations as 2 if these collocations had not any semantic similarity, as 1 if the pair of collocations had some semantic similarity and as 0 if they obviously found it difficult to answer.

Table V shows the values of the average precision of our approach calculated for three types of collocations.

TABLE V
THE CALCULATION OF AVERAGE PRECISION OF OUR APPROACH FOR THREE TYPES OF COLLOCATIONS

Type of collocations	**Average precision**
Substantive	0.781
Attributive	0.644
Verbal	0.627

Such precision is close to results of the other studies [4], [5]. In our opinion, the reason why the data show relatively low results is mistakes of the POS tagging and UD-parser.

In the future research we intend to broaden the scope of the study on semantic equivalence. In particular, there is a need for calculating the recall of our experiment and extending the approach to some other languages. Multilingualism of Wikipedia on the one hand, and the independence of each language version of this encyclopedia on the other, give the opportunity to create models that can help to identify content with the highest quality [15]. Therefore, presented approach can be used to define new metrics for the tasks of the quality texts assessment.

ACKNOWLEDGMENT

This research is supported by the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan (project No. AP05131073 – Methods, models of retrieval and analyses of criminal contained information in semi-structured and unstructured textual arrays).

REFERENCES

[1] C. De Boom, S. V.Canneyt, S. Bohez, T. Demeester, B. Dhoedt, “Learning Semantic Similarity for Very Short Texts,” *Pattern Recognition Letters,* vol. 80, 2016, pp. 150–156. DOI: 10.1109/ICDMW.2015.86

[2] J. Ganitkevitch, B. V. Durme, C. Callison-Burch, “PPDB: The paraphrase database,” *in Proc. of the 2013 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 2013, pp. 758–764.

[3] H. Wu, M. Zhou, “Synonymous Collocation Extraction Using Translation Information,” *in Proc. of the 41st Annu. Meeting on Association for Computational Linguistics,* Stroudsburg, PA, USA, vol. 1, 2003, pp. 120–127. DOI: 10.3115/1075096.1075112

[4] M. Pasca, P. Dienes, “Aligning Needles in a Haystack: Paraphrase Acquisition Across the Web,” *in Proc. of the Second Int. Joint Conf.: Natural Language Processing,* Korea, 2005, pp. 119–130. DOI: 10.1007/11562214_11

[5] R. Barzilay, Kathleen R. McKeown, “Extracting Paraphrases from a Parallel Corpus,” *in Proc. of the 39th Annu.Meeting on Association for Computational Linguistics,* Stroudsburg, PA, USA, 2001, pp. 50–57. DOI: 10.3115/1073012.1073020

[6] J. Ganitkevitch, C. Callison-Burch, C. Napoles, B. V. Durme, “Learning Sentential Paraphrases from Bilingual Parallel Corpora for Text-to-Text Generation,” *in Proc. of the Conf. on Empirical Methods in Natural Language Processing,* 2011, pp. 1168–1179.

[7] B. Dolan, C. Quirk, C. Brockett, “Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources,” *in Proc. of the 20th Int. Conf. on Computational Linguistics,* Geneva, Switzerland, 2004. DOI: 10.3115/1220355.1220406

[8] L. Han, A. Kashyap, T. Finin, J. Mayfield, J. Weese, "UMBC EBIQUITY-CORE: Semantic Textual Similarity Systems", *in Proc. of the Second Joint Conf. on Lexical and Computational Semantics,* vol. 1, 2013, pp. 44–52.

[9] T. Kenter, M. de Rijke, “Short Text Similarity with Word Embeddings, ”*in Proc. of the 24th ACM Int. Conf. on Information and Knowledge Management,* 2015, pp. 1411–1420. DOI: 10.1145/2806416.2806475

[10] W. Lewoniewski, K. Węcel, W. Abramowicz, “Relative Quality and Popularity Evaluation of Multilingual Wikipedia Articles,” *Informatics,* 2017. DOI: 10.3390/informatics4040043

[11] S. Petrasova, N. Khairova, “Automatic Identification of Collocation Similarity,” *in Proc. of 10th Inter. Scientific and Technical Conf.: Computer Science & Information Technologies,* Lviv, 2015, pp. 136–138. DOI: 10.1109/STC-CSIT.2015.7325451

[12] S. Petrasova, N. Khairova, “Using a Technology for Identification of Semantically Connected Text Elements to Determine a Common Information Space,” *Cybernetics and Systems Analysis,* Springer, vol. 53 (1), 2017, pp. 115–124. DOI: 10.1007/s10559-017-9912-z

[13] Joakim Nivre Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, et al., “Universal Dependencies v1: A Multilingual Treebank Collection,” *in Proc. of the Tenth Int. Conf. on Language Resources and Evaluation,* Paris, France, 2016

[14] T. McEnery, A. Hardie, “Corpus Linguistics: Method, Theory and Practice,” Cambridge University Press, 2012.

[15] Lewoniewski W. “Enrichment of Information in Multilingual Wikipedia Based on Quality Analysis,” *Lecture Notes in Business Information Processing,* vol 303. Springer, Cham, 2017, pp 216–227. DOI: 10.1007/978-3-319-69023-0_19