Detecting Collocations Similarity via Logical-Linguistic Model

Nina Khairova, Svitlana Petrasova
National Technical University "Kharkiv Polytechnic Institute",
Kyrpychova str., 61002, Kharkiv, Ukraine
khairova@kpi.kharkov.ua, svetapetrasova@gmail.com

Orken Mamyrbayev
Institute of Information and Computational Technologies,
125, Pushkin str., 050010, Almaty, Republic of Kazakhstan
morkenj@mail.ru

Kuralay Mukhsina
Al-Farabi Kazakh National University,
71 al-Farabi Ave., Almaty, Republic of Kazakhstan,
kuka_ai@mail.ru

Abstract

Semantic similarity between collocations, along with words similarity, is one of the main issues of NLP. In particular, it might be addressed to facilitate the automatic thesaurus generation. In the paper, we consider the logical-linguistic model that allows defining the relation of semantic similarity of collocations via the logical-algebraic equations. We provide the model for English, Ukrainian and Russian text corpora. The implementation for each language is slightly different in the equations of the finite predicates algebra and used linguistic resources. As a dataset for our experiment, we use 5801 pairs of sentences of Microsoft Research Paraphrase Corpus for English and more than 1 000 texts of scientific papers for Russian and Ukrainian.

1 Introduction

Nowadays, linguistic resources are not only a part of any linguistic study but an important base for designing NLP applications such as search engines, machine (-assisted) translation, context-sensitive ads, document clustering, automatic essay scoring, business intelligence (e.g. sentiment analysis) and text summarization. Linguistic resources typically include linguistic ontologies, monolingual and multilingual corpora and various kinds of dictionaries.

Thesauri, where words are associated with semantic relations to each other, are of particular importance among all dictionary types. However, in order to create a thesaurus, lexicographic researches, the analysis of the lexical structure of languages, exploring of the text characteristics and similar labour-intensive studies must be conducted (Jarmasz and Szpakowicz, 2003). The thesaurus design process can be accelerated by the automation of the close concepts identification step.

In a general way, such concepts are represented by a single word, but sometimes a concept can be represented by two or three related words. As of today, a sufficient number of approaches exists to find and extract semantically similar words from a corpus automatically. However, measuring the semantic similarity between word groups or collocations is a more challenging task which has no satisfactory solution to date.

In our study, we propose the logical-linguistic model to identify semantic similarity of collocations. Generally, a collocation is considered as a combination of two lexical units in syntactic and semantic relations that co-occur in the text non-randomly. The probabilistic study of collocation occurrence is

beyond the scope of this research, though. We assume that two-word combinations are considered as collocations if they occur more than once in synonymous meanings.

We created the models for English, Ukrainian and Russian languages. Using these models, in general, allows extracting semantically similar collocations from a text corpus automatically in order to generate a first draft of the thesaurus.

2 Related work

The most explored level of text similarity for different languages is the level of words. In this way, we can distinguish two classes of words similarity algorithms. The first approach is based on the exploitation of a thesaurus (Pirró and Seco, 2008; Pedersen et al., 2007). The second methods and algorithms group of word similarity identification focuses on distributional models of meaning in a corpus (Islam and Inkpen, 2006; Han et al., 2013; Akermi and Faiz, 2012).

There is much less research related to the measurement of similarity between sentences or short text fragments (Islam and Inkpen, 2008). In order to evaluate the degree of two English sentences semantic similarity, Sultan et al. exploited an unsupervised system that relied on word alignment (Sultan et al., 2014) or combined a vector similarity feature with alignment-based similarity (Sultan et al., 2015). Now quite a few researchers apply align words algorithms in order to compute the semantic similarity between two sentences. McCrae et al. (2016) also exploited the idea of creating monolingual alignments to assess the degree of semantic similarity of sentences. However, they proposed to use soft alignment, where they produced a score indicating how likely one word in the sentence was to be aligned to another word in the other sentence.

Dang et al. (2016), like many others, drew on tweets as short text fragments. They proposed to use Wikipedia as an external knowledge source and a corpus-based word semantic relatedness method to determine whether two tweets are semantically similar or not. Rakib et al. (2016) also benefited from an external knowledge source such as Google-n-grams. They computed relatedness strength between two phrases using the sum-ratio technique in conjunction with cosine similarity via bi-gram contexts from Google-n-grams. Recently Boom et al. (2015) used a hybrid method that united word embedding and tf-idf information of a text fragment into a distributed representation of very short text fragments semantically close to each other.

Increasingly, the task of measuring the semantic similarity of short text fragments is being integrated into the common challenges of the paraphrase. However, in general, such researches involve semantic similarity of sentences (Ganitkevitch et al., 2013; Pavlick et al., 2015). Extracting paraphrase fragment pairs, Wang and Callison-Burch (2011) used a comparable corpus, and in the next study they utilized parallel corpora considering discourse information (Regneri and Wang, 2012).

Measuring the semantic similarity of collocations is a more challenging task than searching words or sentences with similar meaning. This is connected to the fact that both identifying collocations and establishing their synonymy must be involved in the process of detecting semantically similar items.

3 The proposed method for detecting semantic similarity

We propose a method to detect and extract semantically similar collocations from text corpora. In our study, we consider semantically similar collocations as synonymous collocations with certain assumption having been made.

The method is based on the logical-linguistic model (Khairova et al., 2015) that: (1) formalizes semantic and grammatical words characteristics of prospective collocations by means of the subject variables; (2) identifies substantive, attributive and verbal collocations by means of equations of the finite predicates algebra; (3) formalizes structures of semantically similar collocations via the logical-algebraic equations. Additionally, we exploit POS-tagging and thesauri as linguistic resources of a particular language. POS-tagging is applied to extract grammatical characteristics of words, and thesauri are applied

to find potential synonyms of the collocation words that were identified in the first and second phases of the model.

Fig. 1 shows the structural scheme of the method, which highlights the synergy between the logical-linguistic model and linguistic resources of a particular language.

We provide the model for extraction of semantically similar collocations from Ukrainian, Russian and English text corpora. The semantic cohesion between 2 words in a collocation is expressed by morphological and syntactic relations in all these languages. The distinctions between the implementations of the model for the various languages are in (1) different values of the subject variables, (2) slightly different logical-linguistic equations of substantive, attributive, verbal collocations and (3) discrepancy of logical-algebraic equations of the semantic similarity for collocations. The main reason for this differentiation is that the semantic cohesion in the Ukrainian and Russian languages is represented by a range of grammatical cases while the order of words and existence of prepositions represent the semantic relations in English.

In this way, the model involves the following steps. The first step is preprocessing when we tag a text corpus. POS-tagging is carried out in order to identify substantive (Noun-Noun), attributive (Adjective-Noun), and verbal (Verb-Noun) collocations. In the next step, we identify characteristics of collocation words. Furthermore, using a thesaurus we get synonymous pairs of words that were found in the previous steps. The last step, we determine pairs of semantically similar collocations using the predicates of equivalence and then find the pairs in a corpus.

In the first preprocessing stage, we perform POS-tagging by means of NLTK Python library to identify two adjacent words as a possible collocation. For example, to identify substantive (Noun-Noun) collocations in the Ukrainian language, we find the main word marked <NN >and one of the other tags, which represents the grammar case, must be <Nom >, <Gen>, <Dat>, <Acc>, <In>or <Pr>. The dependent word of substantive collocations in the Ukrainian language must be marked as a noun too (<NN >). Nevertheless, its case must be marked only as <Gen>.

In this way, substantive collocations in Ukrainian can be defined by the following logical-linguistic equation:

$$(x^{NNom} \vee x^{NGen} \vee x^{NDat} \vee x^{NAcc} \vee x^{NIn} \vee x^{NPr}) y^{NGen} = 1 \tag{1}$$

Similarly, we can determine attributive and verbal collocations by the following logical-linguistic equations respectively:

$$\begin{gathered} y^{ANom} x^{NNom} \vee y^{AGen} x^{NGen} \vee y^{ADat} x^{NDat} \vee y^{AAcc} x^{NAcc} \vee \\ y^{AIn} x^{NAIn} \vee y^{APr} x^{NPr} = 1 \end{gathered} \tag{2}$$

$$x^{VNonRef} y^{NAcc} = 1 \tag{3}$$

In the equations (1)- (3) the subject variable x describes a set of possible grammatical characteristics for a main collocation word and the subject variable y describes a possible set of characteristics for a dependent word of the collocation.

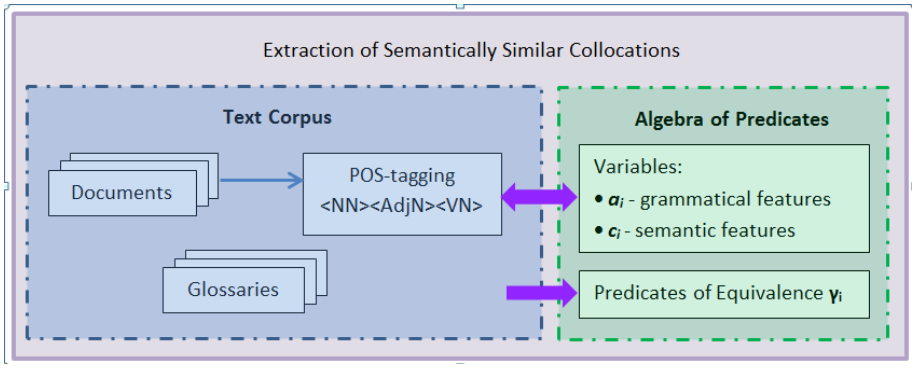

Figure 1: The structural scheme of our method

The next step, we define a set of grammatical and semantic characteristics of words for the Ukrainian, Russian, and English languages using two subject variables that define grammatical (a^i) and semantic (c^i) categories of the language. Every subject variable z^i equals to 1 if main or dependent words might have these i characteristics, and it equals to 0 otherwise. The grammatical characteristics of collocation words are mostly received as a result of POS-tagging.

As illustrated above, for Ukrainian and Russian languages the main grammatical characteristics that show the dependency in collocations are a part of speech, transitivity (in case of verbs) and a case. As for English, such grammatical characteristics, apart from POS and verb transitivity, are the existence of a particular preposition and/or the existence of the apostrophe at the end of the word and/or the existence of any form of the verb "to be" in the phrase and the position of the noun concerning a verb (Khairova et al., 2016).

The subject variable c^i defines 6 semantic cases for all three languages: c^{Ag} – an Agent, c^{Att} – an Attribute, c^{Pac} – a Patient, c^{Adr} – an Addressee, c^{Ins} – an Instrument, c^{M} – a Location or Content.

In our model, a set of possible grammatical and semantic characteristics for the main collocation word is defined by the predicate P(x). The predicate P(y) specifies grammatical and semantic characteristics of the dependent word in collocations. Therefore, we define two-word collocations via the double predicate P(x, y) that combines two previous predicates. For the Ukrainian and Russian languages, the predicate is following:

$$\begin{aligned} P(x,y) =&(a_y^{ANom} \vee a_y^{AGen} \vee a_y^{AAcc} \vee a_y^{ADat} \vee a_y^{AIn} \vee a_y^{APr})(a_x^{ANom} c_x^{Ag} \vee a_x^{NGen} c_x^{Att} \vee \\ & a_x^{NAcc} c_x^{Pac} \vee a_x^{NDat} c_x^{Adr} \vee a_x^{NIn} c_x^{Ins} \vee a_x^{NPr} c_x^{M}) a_y^{NGen} c_y^{Att} \vee a_x^{VNonRef} a_y^{NAcc} c_y^{Pac} \end{aligned} \tag{4}$$

While in the case of English, the predicate that identifies grammatical and semantic characteristics of words in two-word collocations is following:

$$\begin{aligned} P(x,y) =& a_y^{AAtt} a_x^{NSubj} c_x^{Ag} \vee a_x^{NSubj} c_x^{Ag} a_y^{APr} \vee (a_x^{NSubj} c_x^{Ag} \vee a_x^{NSubjOf} c_x^{Ag}) \\ & (a_x^{NObj} c_x^{Att} \vee a_x^{NObjOf} c_x^{Att}) \vee a_x^{VNonRef} a_y^{NObj} c_y^{Pac} \end{aligned} \tag{5}$$

For example, the correlation of semantic and grammatical characteristics of Ukrainian attributive collocations such as "technical facilities" ("tekhnichni zasoby") or "engineering tools" ("inzhenerni instrumenty") satisfies the conjunction $a_y^{ANom} a_x^{NNom} c_x^{Ag}$ of the predicate (4). The English word combinations "form the notion" or "create the view" satisfies the conjunction of the grammatical and semantic characteristics of verbal collocations $a_x^{VNonRef} a_y^{NObj} c_y^{Pac}$ of the predicate (5).

The next step, we obtain predicates of the semantic equivalence of two collocations for the substantive (represented by γ_{1L}), attributive (represented by γ_{2L}), verbal (represented by γ_{3L}) ones. For instance, the predicate of semantic equivalence of substantive collocations in Ukrainian and Russian corpora is defined as γ_{1U}:

$$\gamma_{1U}(x_1, y_1, x_2, y_2) = a_{x1}^{NNom} c_{y1}^{Ag} a_{y1}^{NGen} c_{y1}^{Att} \wedge a_{x2}^{NNom} c_{y2}^{Ag} a_{y2}^{NGen} c_{y2}^{Att} \tag{6}$$

We define the predicate of semantic equivalence of verbal collocations in English corpora as γ_{3E}:

$$\gamma_{3E}(x_1, y_1, x_2, y_2) = a_{x1}^{VNonRef} a_{y1}^{NObj} c_{y1}^{Pac} \wedge a_{x2}^{VNonRef} a_{y2}^{NObj} c_{y2}^{Pac} \tag{7}$$

We use thesauri to establish the synonymy between collocates. In the case of English, we utilize WordNet 3.1.0 of 151 806 unique nouns, verbs and adjectives, that contains synsets in every dictionary entry. For the Ukrainian language, we have developed a thesaurus of about 3 000 unique nouns, verbs and adjectives.

We assume that collocations can be considered as semantically similar if the main word x_1 of the collocation is synonymous with the main word x_2 in the second collocation as well as the dependent word y_1 is synonymous with y_2.

Therefore, collocations can be considered to be semantically close if (1) their grammatical and semantic features satisfy the predicate of equivalence and (2) the words of two collocations are synonymous in pairs. Table 1 shows the examples of three types of synonymous collocations extracted from our Ukrainian, Russian and English text corpora.

Table 1: The examples of three types of synonymous collocations extracted from Ukrainian, Russian and English text corpora

Collocations type	**English language**	**Ukrainian, Russian languages**
substantive (Noun-Noun)	health department – health officials	zastosuvannya komputera (the computer application) – vykorystannya noutbuka (the use of a laptop)
attributive (Adjective-Noun)	federal agents – federal investigators	suchasniy metod (the up-to-date method) – inovatsiyniy sposib (an innovative way)
verbal (Verb-Noun)	deliver assessments – present assessments	prepodnosit informatsiyu (to present the information) – predstavlat svedenia (to present the data)

4 Source data and experimental results

To evaluate the effectiveness of the proposed logical-linguistic model, we designed a corpus of more than 1 000 Ukrainian and Russian texts of scientific papers that contain more than 3,5 million words and about 2200 unique words. All papers are devoted to the broad theme of information technologies. As a result of the experiment, we extracted 62738 substantive, 46808 attributive and 3965 verbal semantically similar collocations. These collocations are similar in one or more pairs.

In order to evaluate our experimental results for the Ukrainian and Russian languages, we used experts' opinion. About 500 synonymous pairs of collocations were randomly extracted from the lists of these pairs for each language and presented for judgment. Three experts were asked to compare the similarity of meaning of the collocation pairs on the scale of from 0 to 2: 0 – the collocations don't have any semantic similarity, 2 – the pair of collocations has some semantic similarity, 1 – the experts find it difficult to answer. We considered the collocations in the pair as semantically similar when the average score of experts was more than 1.4. For example, when all the experts rated a pair of collocations as 2, the inter-rater agreement equaled to 2. If collocations were rated by two of experts as 2 and by one expert as 1, the inter-rater agreement equaled to 1.7. However, in cases of the inter-rater agreement of less than 1.4, the pair of collocations is thought as not semantically similar.

To evaluate the effectiveness of the model for extracting semantically similar collocations from the English corpus, we exploit Microsoft Research Paraphrase Corpus (MRPC), which consists of 5801 pairs of sentences obtained from thousands of news sources on the web. Fig 2 shows the example of the extraction of semantically similar collocations from two semantically similar sentences of the corpus.

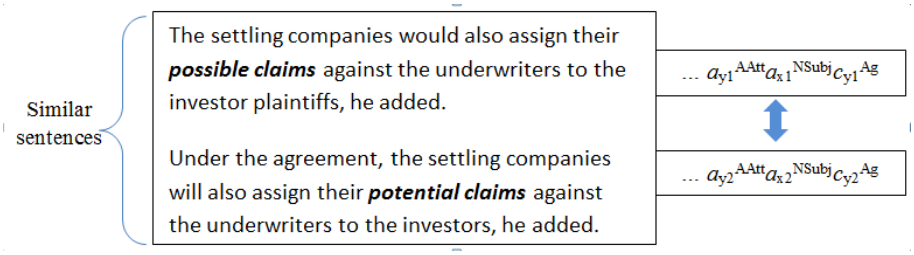

Figure 2: The example of the extraction of semantically similar collocations from two semantically similar sentences of MRPC corpus via the logical-linguistic equations

In MRPC all the pairs of sentences were rated by 2 judges as "semantically equivalent" or "non-equivalent". The inter-rater agreement was averaging 83% (Quirk et al., 2004).

As a result, precision, showing the correctness of the semantic similarity relation of collocations, is 0.7459 for Ukrainian and Russian texts and 0.8898 for English Microsoft Research Paraphrase Corpus. Relevant approaches extracted paraphrase fragment pairs with the precision of 67%, manually annotating fragment pairs as paraphrases, related or invalid (Wang and Callison-Burch, 2011) and 84%, rating fragment pairs as paraphrases, related or irrelevant with the inter-annotator agreement according to Cohen's Kappa of 0.67 (Regneri and Wang, 2012).

Additionally, using Microsoft Research Paraphrase Corpus we have been able to calculate the recall of the model. To do that we had hypothesized that whether sentences have a similar meaning they must contain similar collocations. Knowing the total amount of similar sentences in the corpus, we assume that each similar sentence pair contains one synonymous collocation pair. Consequently, to evaluate the recall of our experiments, we have computed the ratio between the number of semantically similar collocation pairs found (3650) to the total amount of sentence pairs specified as semantically equivalent (3900). Based on the hypothesis we calculated the recall of our model for English text as 94%.

5 Conclusions and Future Work

This paper proposes a novel logical-linguistic model for extraction of semantically similar two-word collocations from the Ukrainian, Russian and English corpora as an additional option of the first stage of generating the thesaurus automatically.

In order to assess our model, the corpora in various languages are exploited. We compute the precision of the model for Russian and Ukrainian languages on the basis of the corpus that comprises more than 1000 scientific articles devoted to the information technologies themes. To compute the precision of the model for English, we exploit MRPC. Additionally, since the corpus preliminary annotated we are able to calculate the recall of the model.

Our model achieves as a result over 74% precision of extraction of semantically similar collocations from Ukrainian and Russian corpora, about 89% from English one. Moreover, the recall of semantically similar collocations extraction from English Microsoft Research Paraphrase Corpus achieves over 94%. The task for further work is verification of our research results via probabilistic computation of occurrence of synonymous collocations in text corpora.

In future studies, we intend to broaden the scope of collocation types examination and to consider the combination of main parts of speech with auxiliary ones (e.g. prepositions, conjunctions etc.) that go beyond the scope of the model now. Additionally, in prospect, we intend to spread our dataset for free access to carry out similar approaches.

6 Acknowledgment

The research was funded in part 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

Akermi, I. and R. Faiz (2012). A novel method for word-pair similarity computing. *International Journal of Computational Linguistics Research 3*(4), 131–142.

Boom, C. D., S. V. Canneyt, S. Bohez, T. Demeester, and B. Dhoedt (2015). Learning semantic similarity for very short texts. *2015 ieee international conference on data mining workshop (icdmw)*, 1229–1234.

Dang, A., R. Makki, A. Moh'd, A. Islam, V. Keselj, and E. Milios (2016). Real time filtering of tweets using wikipedia concepts and google tri-gram semantic relatedness. *The Twenty-Fourth Text REtrieval Conference Proceedings* (2).

Ganitkevitch, J., B. V. Durme, and C. Callison-Burch (2013). Ppdb: The paraphrase database. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 758–764.

Han, L., T. Finin, P. McNamee, A. Joshi, and Y. Yesha (2013). Improving word similarity by augmenting pmi with estimates of word polysemy. *IEEE Transactions on Knowledge and Data Engineering 25*(6), 1307–1322.

Islam, A. and D. Inkpen (2006). Second order co-occurrence pmi for determining the semantic similarity of words. *LREC*, 1033–1038.

Islam, A. and D. Inkpen (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD) 2*(2), 10:1–10:25.

Jarmasz, M. and S. Szpakowicz (2003). Thesaurus and semantic similarity. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*, 212–219.

Khairova, N., S. Petrasova, and P. S. Gautam, A. (2015). The logic and linguistic model for automatic extraction of collocation similarity. *Econtechmod an international quarterly journal 4*(4), 42–48.

Khairova, N., S. Petrasova, and P. S. Gautam, A. (2016). The logical-linguistic model of fact extraction from english texts. *International Conference on Information and Software Technologies*, 625–635.

McCrae, J. P., K. Asooja, N. Aggarwal, and P. Buitelaar (2016). Nuig-unlp at semeval-2016 task 1: Soft alignment and deep learning for semantic textual similarity. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 712–717.

Pavlick, E., J. Bos, M. Nissim, C. Beller, B. V. Durme, and C. Callison-Burch (2015). Adding semantics to data-driven paraphrasing. *roc. of ACL-IJCNLP. Beijing, China*.

Pedersen, T., S. Pakhomov, S. Patwardhan, and G. Chute, C. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics 40*(3), 288–299.

Pirró, G. and N. Seco (2008). Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content. *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, 1271–1288.

Quirk, C., C. Brockett, and W. B. Dolan (2004). Monolingual machine translation for paraphrase generation. *In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 142–149.

Rakib, M. R. H., A. Islam, and E. Milios (2016). f: Phrase relatedness function using overlapping bi-gram context. *Canadian Conference on Artificial Intelligence*, 137–149.

Regneri, M. and R. Wang (2012). Using discourse information for paraphrase extraction. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 916–927.

Sultan, M. A., S. Bethard, and T. Sumner (2014). Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics 2*, 219–230.

Sultan, M. A., S. Bethard, and T. Sumner (2015). Dls@ cu: Sentence similarity from word alignment and semantic vector composition. *SemEval*, 148–153.

Wang, R. and C. Callison-Burch (2011). Paraphrase fragment extraction from monolingual comparable corpora. *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, 525–60.