Computational theory of semantics representation in scientific and technical texts

Y M Vishnyakov, R Y Vishnyakov

1 Mathematics and Computer Science Department, Kuban State University, 3500401, Krasnodar, Russia
E-mail: Jury.Vishnyakov@gmail.com

Abstract. The work, described in this paper, continues the authors’ research in the field of computational semantics of natural languages. The main focus of the paper is on the words processing in the sentence. Also the paper considers computational theory of the semantic interpretation of natural language texts. Furthermore, the computational model of a semantics representation in the form of the semantic schemes for the Russian language is discussed. The provided methodology is the constructed rules to compare semantic proximity of the semantic schemes. The obtained results show a qualitative difference between the proposed computational semantic model and the traditional model based on the frequency paradigm is shown. Finally, the provided examples of the semantic comparison demonstrated that the proposed computational semantic model is able to outperform models based on the frequency paradigm.

1. Introduction

Nowadays, the processing of natural language (NLP) texts in such areas as information extraction, information retrieval, classification of documents and etc., is based on the frequency paradigm [1–3], of the semantics model. Its meaning is reduced to determining the relevance of the text by frequency characteristics of words, and thus in fact the frequency of occurrence of words in the texts acts as a model of the semantic component of the text. At the same time, the efficiency of information processing in such a paradigm actually had reached the limit and can no longer satisfy users [4, 5]. Furthermore, the current situation is aggravated by an avalanche-like growth of the information volume and by increasing demand of the quality of the processing itself. All this taken together strongly dictates the need of new solutions in the modeling of semantics outside the frequency paradigm and strengths attention to other ways of modeling semantics.

Due to the ambiguity (polymorphism) of the natural language and the absence of unambiguous relations between the syntagmatic and semantic structures [6] of texts the problem turns into a category that is difficult to formalize. Though, to day, no general suitable solution has been found in the processing of semantics. Nevertheless, in particular language areas were obtained quite good results. This problem is attracting a lot of attention from the research community, and is being widely discussed at scientific meetings of all levels [7–11].
The studies covered in this paper are oriented on Russian texts, written in scientific and/or technical style. A distinctive feature of those texts is that they fit into the logic of plausible reasoning. As a consequence, the sentences of such texts approach semantic accuracy and uniqueness. Moreover, this distinctive feature allows applying formal methods for semantic interpretation. Since such sentences lacking emotional coloring, the coupling of semantics to the syntactic structure of the text is becoming more rigid.

The proposed work is a part of our research on the semantic interpretation of texts of the scientific and technical style. This paper is considered as a review and is based on the materials of the authors, in particular [9–11].

2. Methods

A word in natural language is considered on two levels. First level is a so-called "sign", which exists as an object (in a written or pronounced way) and is an element of a speech. Second level is the set of meanings designated by a word and a lexical meaning of the word. Thus, this set of meanings exists in the context of the subjects and may vary from one subject to another one. Same time not all sets of meanings of a word are used in a particular text, but only one or some of their subsets according to the meaning of the text. Identification of the sets of meanings between two subjects in the text is based on the dependencies of words in the text and on the fundamental concept of natural language such as word-combination.

A word-combination is a pair of words, where only one is the main word, and the other is a dependent one. A dependent word always clarifies the meaning of the main word and plays a role of a context. We could formalize this case as follows. Let $q$ be a chain of words $a_i$, see equation 1, which is an entire textual fragment (TF).

$$q = a_1 a_2 ... a_n$$  \hspace{1cm} (1)

Integrity means that all words in the $q$ chain are included in the word combinations, forming a direct subordination relation of the set of words in TF. This relation is of one-to-one type. If $a$ in the word combination is the main word and $b$ is a dependent one, then this dependence will be represented as follows, where the arrow indicates the direction of the dependence of the words:

$$\overrightarrow{a} : b.$$  \hspace{1cm} (2)

If the word $a$ is the main word in a few word combinations with the dependent words $b_1, b_2, ... , b_p$, then this is represented by the following entry:

$$a : \{b_1, b_2, ..., b_p\}.$$  \hspace{1cm} (3)

The record (3) is a bunch of contextual words of $a$.

We denote by $S(a_i)$ the set of semantic values of the word $a_i$ from $q$ (1). Thus, the set of semantic values of TF (1) can be represented in the form of some function:

$$S(q) = (S(a_1), S(a_2), ..., S(a_n)).$$  \hspace{1cm} (4)

Now we will expand this function. From the above, it is obvious that the meaning of a word combination is a subset of the meanings of the main word. The subset is defined by the meaning of the dependent word that forms the context in the phrase. We define this in the form of the following expression:

$$S(\overrightarrow{a} : b) = S(a)|_{S(b)}.$$  \hspace{1cm} (5)

For (5) is true the following:

$$S(\overrightarrow{a} : b) \subseteq S(a); S(\overrightarrow{a} : b) \neq S(\overrightarrow{b} : a); S(\overrightarrow{a} : \overrightarrow{a}) = S(a).$$  \hspace{1cm} (6)
Taking (5) into account, we determine the meaning of the entire piece of text (1). Then, if the word $a_i$ is the main in the $q$ of text, then the following relation is always valid:

$$S(q) \subset S(a_1),$$

Consequently, the meaning of the piece of the text is a subset of the set of meanings of its main word. Let’s consider the word combination. To do so, we define the operation of contextual refinement of the meaning of the main word by dependent one as follows:

$$S(a : b) = S(a) \cap S(b),$$

Here $\cap$ — an operation of refinement of the contextual meaning, where the arrow above determines the direction of the dependence of words in the word combination. Applying (8) to the context link (3), we obtain the following:

$$S(a : (b_1, ..., b_p)) = S(a : b_1) \cap S(a : b_2) \cap ... \cap S(a : b_p),$$

or otherwise:

$$S(a : (b_1, ..., b_p)) = \cap_p S(a : b_i).$$

Here, the sign $\cap$ — the operation of intersection.

By applying (8) — (10) it is possible to expand an expression of functional for the piece of the text (4). Below is an example of the text fragment "international recognition of educational programs of Russian universities" in Russian language:

$$q = "\text{международное признание образовательных программ российских вузов"},$$

For the convenience, we encode the words in the example expression by Latin letters. Further, we will expand the functional expression for the given TF $q$, which will have the following form:

$$S(q) = (S(a) \cap S(b)) \cap (S(b) \cap (S(d) \cap S(e) \cap (S(e) \cap S(f)))).$$

Supplementary, taking into account the introduced operation of the contextual refinement of the meaning for the given expression, it is possible to design a computational algorithm. Let’s consider the task to compare texts for semantic proximity. The obvious way to solve this task is to directly compare the results of the functional of the meaning of the texts. However, a number of difficulties arise here such as a requirement to define sets of semantic meanings of words. Another approach is based on the indirect comparison of the computational procedures since it does not require the specification of the semantic meanings of words, assuming that they are apriori identical. However, the second approach requires that the structures of computational procedures should be brought to some identical comparison field/dimension.

Studies have shown that an indirect way of comparison has a number of additional advantages. Thus, a further exposition will be dedicated to the indirect comparison.

In mathematics, there is a specific notation to record mathematical expressions — the reverse Polish notation (RPN). The RPN does not contain parentheses and its calculations are performed at once from left to right. The operations are performed as they occur during the passage of the expression. Therefore, the structure of the RPN [13] uniquely determines the structure of the computational procedure. We use these features and construct the representation of the functional of the meaning in the RPN. We omit specific discussions; we provide directly the final results.
In the notation of the RPN, the operation of contextual refinement of the meaning for the words combination (8) will look like:

\[ S(\overrightarrow{a : b}) = S(b) S(a) \cap . \]  

(12)

In (12), compare to (8), to preserve the dependence of words, the arrow over the operation changes its direction. Also, the RPN of the contextual connection (9) with consideration of (12) takes the following form:

\[ S(\overrightarrow{a : (b_1, \ldots, b_p)}) = S(b_1) S(a) \cap \ldots \cap S(b_p) S(a) \cap \cap . \]  

(13)

Here \( p \) corresponds to the arity of the intersection operation.

Applying (11) and (13) to the functional of the meaning of the example of TF (11), we obtain its RPN of the following form:

\[ S(q) = S(a) S(b) \cap \cap S(c) S(d) \cap \cap S(f) S(e) \cap \cap S(d) \cap \cap S(b) \cap \cap \]  

(14)

The steps of the RPN calculations (14) are as follows:

(i) \( S(q) = S(a) S(b) \cap \cap S(c) S(d) \cap \cap S(f) S(e) \cap \cap S(d) \cap \cap S(b) \cap \cap \)

(ii) \( S(q) = r_1 S(c) S(d) \cap \cap S(f) S(e) \cap \cap S(d) \cap \cap S(b) \cap \cap \)

(iii) \( S(q) = r_1 r_2 S(f) S(e) \cap \cap S(d) \cap \cap S(b) \cap \cap \)

(iv) \( S(q) = r_1 r_2 r_3 S(d) \cap \cap S(b) \cap \cap \)

(v) \( S(q) = r_1 r_5 S(b) \cap \cap \)

(vi) \( S(q) = r_1 r_6 \cap \cap \)

(vii) \( S(q) = r_1 r_5 S(b) \cap \cap \)

Here the variables \( r_i \) intend to store the intermediate results of each step. Thus, the transformation of the functional of the meaning into the notation of the RPN uniquely determines its computational procedure.

Next, we construct a graphical representation of the computational procedure. To do this, we assign to each operation a graphic element in which the rectangle represents the operation, the round vertices on the left are input data, on the right are the output data (result). Such a graphical representation of an operation will be called an element of meaning. Combining the elements of the meaning of expression (14) in the order given by the computational procedure, we obtain a graphical representation of the entire computational procedure. This representation we will call the semantic scheme. The semantic scheme for the example of a textual fragment (TF) (14) is presented in figure 1:
3. Results and Conclusion

In this section we discuss how to compare semantic proximity of TF. Suppose we are given a comparison sample $q$ of the TF in the form of (1). This sample is comparable to TF $t$ by semantic proximity and has a form of:

$$t = b_1 b_2 \ldots b_m.$$  \hfill (15)

We introduce the criterion of semantic proximity (CSP) as follows:

$$C_{prox}(S(q), S(t)) \in D;$$

$$D = [0..1].$$  \hfill (16)

Here, $C_{prox}$ is a CSP, $S(q)$ — the set of semantic values of the sample, $S(t)$ — the set of semantic values of $t$ and $D$ is the interval of the CSP values of $C_{prox}$. If $C_{prox} = 0$, then $q$ and $t$ are semantically not close. If $C_{prox} = 1$ there is a complete semantic match.

We note that in the general case the following relation is always valid:

$$C_{prox}(S(q), S(t)) \neq C_{prox}(S(t), S(q)).$$  \hfill (17)

Let us represent the CSP in the coinciding form by meaning elements of fraction in the semantic scheme of the sample $q$ and the compared TF $t$ to the total number of the sample elements of the meaning:

$$C_{prox}(S(q), S(t)) = p/m.$$  \hfill (18)

Here $m$ is the number of elements of the meaning in the semantic scheme of $q$ sample, $p$ is the number of matching elements of the $q$ meaning in semantic scheme of $t$ sample respectively. Further, the frequency criterion of relevance $\phi$ is a special case (18), when $m$ represents the number of words in the TF $q$, and $p$ is the number of words from $q$ that coincide with the words in $t$. Let us explain the semantic comparison of the sample $q$ and the examples of specific textual fragments. We compare the CSP and the frequency criterion of relevance (examples are taken from [12]). Further, for the illustration and understanding purpose, Russian words of the sample $q$ are inscribed in the figures of the semantic schemes. They are representing the meaning over the vertices of the rounds. First example is a piece of text in Russian in the form of a sentence:
Правительство России проводит политику, направленную на международное признание образовательных программ российских вузов.

The sentence in the example is equivalent to the sentence in English: The Russian government pursues a policy aimed at the international recognition of educational programs of Russian universities.

In this example the sample $q$ is a substring of a sentence (italicized) and its semantic schema completely fits into the semantic scheme of the sentence. Thus, $C_{prox}(S(q), S(t)) = 1$ and $\phi = 1$.

Second example represents another TF: Образовательные программы российских вузов нуждаются в международном признании. The same sentence in English is Educational programs of Russian universities need international recognition.

In the second example, by underlined italicized substrings, we show the matched fragments of the text with the first example.

Now, we denote the coinciding elements of the meaning on the semantic schema of the sample by bold lines so that the pattern of the sample will take the following form represented in figure 2.

Figure 2. The comparison result of the proximity of the TF $q$ and the second example.

The semantic scheme of the sample contains 5 coinciding elements of meaning, and therefore its value $C_{prox}(S(q), S(t)) = 5/7$ is still quite large. At the same time, the proximity criterion $\phi = 1$. According to this criterion there is no change of the meaning, since all words of the sample are in the text of the second example.

In the third example is the following sentence: Международное признание образовательных программ поднимает репутационный рейтинг российских вузов.

The equivalent in English: International recognition of educational programs raises the reputation of Russian universities.

While comparing the semantic diagram of the TF shown on figure 3 and the textual fragment $q$, it becomes obvious that the semantic schemes of the TF $q$ and the third example have only three coincident elements of meaning, therefore $C_{prox}(S(q), S(t)) = 3/7$. Indeed, the semantic proximity of the TF $q$ and the third example precisely estimate that fact. However, $\phi = 1$ and the frequency criterion again is not sensitive to the change of the meaning.
Fourth example is the following sentence: **Международное признание вузов увеличивает возможности России по привлечению иностранных студентов.** English version of the sentence — **International recognition of universities enhances Russia’s ability to attract foreign students.**

The result of comparison of the TF q and the text of fourth example is shown in figure 4.

Figure 4 shows that the semantic diagrams of the TF and text of fourth example have only one matching element of meaning and the criterion is sufficiently small. At the same time, the value of the frequency criterion $\phi = 0.5$ is still high, as half of the words in the TF intersects with the text of the example.

The text of the fifth example is as follows: **Международные программы академических обменов и участие в них российских ученых способствуют улучшению образовательного процесса вузов России.** The English version — **International programs of academic exchanges and participation of Russian scientists in them contribute to improving the educational process of Russian universities.**
Figure 5. The comparison result of the proximity of the TF $q$ and the fifth example.

The result of comparison of the TF $q$ and the text of fifth example is represented in figure 5: The semantic schema of the TF does not have the same elements of the meaning as in the text of the fifth example scheme, therefore $C_{prox}(S(q), S(t)) = 0$. At the same time, $\phi = 0.83$, since 5 words of 6 of the TF are in the text of the fifth example.

The examples given above, show that the semantic proximity criterion $C_{prox}$ really reflects the semantic proximity of the TF and a text, because it takes into account not only the number of matching words in the compared text fragments, but also their connection. At the same time, the frequency criterion $\phi$, which takes into account only the number of matching words, does not always reflect the real semantic proximity of the compared TF. This is particularly evident in the fifth example.

Further, the procedures for determining the proximity of textual documents are constructed. This is done on the basis of the semantic criterion of the proximity of the texts. Thus, the work described in [12] is proposing a solution based on the concept of a fuzzy scale and a linguistic variable.

Thus, the work covers the approach of the semantic interpretation of the text in the scientific and technical style on the basis of a computational procedure. Also, the concept of the semantic scheme and the semantic proximity criterion is constructed. Furthermore, we have demonstrated examples how to compare the TFs to semantic proximity. The computational model of the semantic interpretation is formal and thus allows to retrieve a highly-precised information or to classify documents.

4. Acknowledgements
The work was carried out within the framework of the grant: 16-07-00241_17 of Russian Foundation for Basic Research.

References
[1] Feldman R, Sanger J, 2007, The text mining handbook: advanced approaches in analyzing unstructured data Cambridge university press
[2] Delgado M, Martín-Bautista M J, Sánchez D, Vila M A, 2000, Mining text data: special features and patterns Pattern detection and discovery,140–153, Springer.
[3] Tseng Y-H, Lin C-J, Lin Y-I, 2007, Text mining techniques for patent analysis Information Processing & Management, 43(5):1216–1247, Elsevier.
[4] Jing L-P, Huang H-K, Shi H-B, 2002, Improved feature selection approach TFIDF in text mining *Machine Learning and Cybernetics, 2002. Proceedings. 2002 Internationa*, 2:944–946, IEEE.

[5] Hu K, Wu H, Qi K et Al., 2017, A domain keyword analysis approach extending Term Frequency-Key word Active Index with Google Word2Vec model *Scientometrics*, 1–38, Springer.

[6] Cruse A, 2011, *Meaning in language: An introduction to semantics and pragmatics* Oxford University Press UK.

[7] Dobrov B V, Loukachevitch N V, 2011, Multiple Evidence for Term Extraction in Broad Domains. *RANLP*, 710–715.

[8] Efimenko I V, 2006, Obrabotka yestestvennyazvykovykh tekstov: ontologichnost’ v lingvistike i diskursivnost’ v izvechenii znanii (Processing of texts: ontology in linguistics and discourse in the extraction of knowledge) URL: http://www. rai. org/resurs/papers/kii-2006/doklad/Efimenko_2.doc.

[9] Vishnyakov R, Vishnyakov Y, 2014, Representation of semantically cohesive sentence fragments in scientific and technical texts *Applied Machine Intelligence and Informatics (SAMI), 2014 IEEE 12th International Symposium on*, 295–298, IEEE.

[10] Vishnyakov Y, Vishnyakov R, 2013, The Linguistic Proximity in Information Retrieval and Document Classification. *14th IEEE International Symposium on Computational Intelligence and Informatics to be held on November 19-21*.

[11] Vishnyakov Y M, Vishnyakov Y Y 2016, Vychislitel’naya semanticheskaya interpretatsiya tekstov nauchno-tehnicheskogo stilya (Computational semantic interpretation of texts of scientific and technical style) *Sozvremenyye naukovenkiye tehnologii,236–242,Obshchestvo s ogranichennoy otvetstvennost’yu" Izdatel’skiy Dom" Akademiyi Yeستestvoznaniya".*

[12] Vishnyakov Y M, Vishnyakov Y Y 2016, Vychislitel’naya semanticheskaya interpretatsiya tekstov nauchno-tehnicheskogo stilya (Computational semantic interpretation of texts of scientific and technical style) *Sozvremenyye naukovenkiye tehnologii,236–242,Obshchestvo s ogranichennoy otvetstvennost’yu" Izdatel’skiy Dom" Akademiyi Yeستestvoznaniya".*

[13] Koncel-Kedziorski R, Hajishirzi H, Sabharwal A et Al., 2015, Parsing algebraic word problems into equations *Transactions of the Association for Computational Linguistics, 3:585–597.*