Measurement of semantic proximity within computational theory of semantic interpretation

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

Abstract. This work develops a computational theory of the semantic interpretation of texts in terms of measuring their semantic proximity. For this, the notion of semantic proximity is determined on the basis of a linguistic variable. For a linguistic variable, a transition is proposed from the values of the coefficients of semantic proximity of textual fragments to fuzzy variables, forming a term-set of its values. To calculate the values of the linguistic variable, consistent decision rules were proposed and constructed. On the basis of these rules, a procedure for the formation of an information output was constructed. This procedure is applied for further information retrieval. The proposed solutions allows to organize an exact information search based on the measurement of the semantic proximity of documents to the information request.

1. Introduction
To date, in the field of information retrieval, it is necessary to determine the semantic proximity or, in other words, the relevance of returned documents to the initial query. Obviously, measurement methodologies and a model of semantic proximity are the key in solving the problems In the classification or categorization tasks. In retrospective, it should be noted that in a semantic model used in real-case scenario such as information retrieval, is used the paradigm of frequency. The main idea of such paradigm can be expressed in the following obvious statement: the more often the words of the query are found in the text of the document the more this document is relevant to the query. Thus, the meaning of semantic proximity is actually reduced to the frequency of query words occurrence in the texts of documents. Further, the frequency characteristics of occurrence are used as the basis of a semantic model. Also, similar “frequency” is used in the classification of documents. However, the frequency model of the relevance does not satisfy users anymore. We should notice that such models based on frequency is showing bad performance in distinguishing texts that use same terminology. The situation is aggravated by the avalanche-like growth of information volumes and the increasing demands on the accuracy of the processing itself. All this taken together raised the urgent question of finding new solutions, increasing attention to other ways of modeling semantics. Due to the polymorphic nature of languages and the ambiguous relationship between the syntagmatic and semantic structures of texts, the problem is transferred to the category of difficult to be formulated tasks. Nevertheless, to date, there have been no general solutions in the processing of semantics. However, in particular language areas, the results turned out to be quite acceptable.
This is true for semantic texts, the typical representatives of which are texts of scientific and technical style. The logic of plausible reasoning in them causes their aggravation to semantic accuracy, so the relation of semantics with syntagmatics to unambiguity. The problem itself is widely discussed in scientific community [1–13].

The proposed work represents a part of our research on the semantic interpretation of texts of a scientific and technical style. It is aimed at forming criteria and procedures for comparing texts of documents. The performance is based on the materials of the authors’ works, particularly [9–13].

2. Methods

The computational theory of semantic interpretation is based on the following propositions: if a certain integral text fragment (a chain of words) is defined and there is a certain word in it, which is the main word in the combinations of the dependent words, then this occasion is represented by a special formula expression, a contextual binding. In order to represent the contextual binding in a phrase, an operation of contextual clarification of the meaning is introduced. Since the meaning of a textual fragment is a subset of the meanings of its main word, then these concepts, as well as operations on sets allow us to form a representation of the functional meaning of the text in two notations: 1) in the form of a formula and 2) a dependency tree. However, it is not possible to use these representations to develop simple computational procedures, since there are the same difficulties as in the case of the usual binary notation of operations, causing an ambiguous order of calculation of a formula expression.

The analysis of the formulaic notation of sense functional found that by introducing some additional constraints on the notation it can be reduced to some representation similar to reverse Polish notation (RPN). The main advantage of the RPN is the absence of parentheses and the order of calculation is from left to right. Such a computational procedure can be represented by some oriented graph of a special type, which we denoted as the semantic scheme [9–13].

The study of the properties of the semantic scheme showed that such scheme is the only one for the given representation of the functional of meaning in the reverse Polish record. Therefore, a comparison of the semantic proximity of fragments of texts can be organized indirectly by going to a comparison of their semantic schemes. To compare by semantic scheme, the concept of the meaning element is introduced. It is proposed to make a comparison by counting the same elements of meaning in the semantic schemes of compared text fragments. For example, there are two text fragments: 1) the comparison pattern \( q \) and 2) the fragment \( t \) to be compared with.

Moreover, the semantic pattern of the sample contains \( n \) elements of meaning. The coefficient of semantic proximity (CSP) of the fragment \( t \) to the sample \( q \) is the ratio \( m/n \), where \( m \) is the number of meaning elements of the fragment \( t \) semantic scheme, which are also represented in the semantic scheme of the sample \( q \).

Let \( q \) be a string of some words \( a_i \) and to represent a complete text fragment \( q = a_1,a_2,\ldots, a_i \). If \( a \) is the main word in several phrases with \( b_1, b_2,\ldots, b_p \) dependent words, then this dependence is represented by the expression (contextual binding) of the form \( a:b_1, b_2,\ldots, b_p \). On the phrase, an operation of contextual clarification of the meaning of the main word is defined in the form of an expression: 

\[
S(a:b) = S(a) \bigcap S(b)
\]

where \( \bigcap \) is the operation of contextual clarification of the meaning, and an arrow above it, sets the direction for the dependence of words in the phrase. It is summarized in the contextual linkage by

\[
S(a : (b_1, \ldots, b_p)) = S(a : b_1) \bigcap S(a : b_2) \bigcap \ldots \bigcap S(a : b_p).
\]

So, the operation of contextual clarification of the meaning is represented in RPN as

\[
S(a : b) = S(b) S(a) \bigcap
\]

Here, to preserve the dependence of words, the arrow on the operation changed its direction. The generalization of the concept of RPN on the contextual binding has the following
form: \[ S(a : (b_1, \ldots, b_p)) = S(b_1)S(a) \bigcap S(b_2)S(a) \bigcap \ldots \bigcap S(b_p)S(a) \bigcap p \bigcup, \]

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

Since the document for which its relevance is determined for the sample (query) \( q \) is represented by a set of text fragments (sentences), the comparisons should be made across all text fragments of the document. At the same time each comparison requires to calculate its own CSP. Obviously, the CSP values from a fragment to a fragment will differ. Then, to establish the degree of semantic proximity of a document to a sample, it is necessary to formulate some integral index of semantic relevance. Probably, it would be possible in the simplest case to calculate the average value of the CSP for the whole document and take it for its semantic relevance. However, in this case, some features of relevance associated with its fine tuning will be lost.

The paper proposes a different approach to determine semantic relevance. It is associated with the representation of semantic relevance as values of a linguistic variable of fuzzy logic. Such a representation will express the relevance in the verbal form familiar to the user, and carry out the adjustment as well as training of the method. Indeed, for a linguistic variable, the procedure to calculate its values is constructed once, which makes the procedure for calculating semantic proximity universal. At the same time, the method is adjusted by values setting of the linguistic variable, which are fuzzy variables and values.

Without loss of generality, we provide a concrete example. For the CSP values, we introduce the D scale and select the intervals on it: \([0..p1), [p1..p2), [p2..p3), [p3..1]\). We introduce the linguistic variable \( \text{SemProx} \) and call it semantic proximity. Let us define the term set \( T \) of its values for the linguistic variable \( \text{SemProx} \). To each element of a term set \( T \) we assign the correspondence interval of the scale \( D \). Then we name these intervals with the corresponding types of proximity of the query \( q \) and a fragment of the text \( t \). Now the term set \( T \) can be written as follows: \( T = \{ \text{"weak"}, \text{"relatively weak"}, \text{"strong enough"}, \text{"strong"} \} \). Obviously, each element of the term-set \( T \) is a fuzzy variable of the form \( ((Cprox(S(q),S(t))), \mu(Cprox(S(q),S(t)))) \) where \( \mu \) is the membership function. In the process of passing through the document, the sample \( q \) is sequentially compared to the semantic proximity with each textual fragment of the document and in each comparison is determined by the CSP. For each CSP value is determined the belonging to one of the intervals of the \( D \) scale, and then for each interval the number of such comparisons \( f \) is calculated. These numbers compose the characteristic vector of the document type: \( <f_{\text{weak}}, f_{\text{rel:weak}}, f_{\text{strong:enough}}, f_{\text{strong}} > \). To calculate the proximity of the document to the query \( q \), we construct the following decision rules:

\[
\begin{align*}
\text{IF } (f_{\text{weak}} >& f_{\text{str:en.}} &+ f_{\text{str}}) &\& (f_{\text{weak}} > f_{\text{rel:weak}}) \text{ THEN SemProx=\"Weak";} \\
\text{IF } ((f_{\text{rel:weak}} > (f_{\text{str:en.}} &+ f_{\text{str}})) &\& (f_{\text{rel:weak}}> f_{\text{weak}})) \text{ THEN SemProx=\"Relative";} \\
\text{IF } ((f_{\text{str:en.}} > (f_{\text{weak}} &+ f_{\text{rel:weak}})) &\& (f_{\text{strong:en.}} > f_{\text{str}})) \text{ THEN SemProx=\"Sufficient";} \\
\text{IF } ((f_{\text{str:en.}} > (f_{\text{weak}} &+ f_{\text{rel:weak}})) &\& (f_{\text{str:en.}} > f_{\text{str}})) \text{ THEN SemProx=\"Strong".}
\end{align*}
\]

These rules determine the unambiguous decision about the proximity of the document to the request based on the characteristic vector. Assume the checking of the document based on decision rules as a following procedure: \( \text{Semantic_proximity (D, SemProx)} \). Here, \( D \) represents the document being analyzed, and \( \text{SemProx} \) is the return value of linguistic proximity. The algorithm that implements the inclusion of the document to the final output will be as follows:

\[
\begin{align*}
\text{BEGIN} \\
\text{WHILE } (\text{documents} != 0) \text{ DO} \\
\text{Semantic_proximity (D, SemProx);} \\
\text{IF } (\text{Semprox}=(\text{condition of implication})) \text{ then D } \rightarrow \text{ R} \\
\text{END}
\end{align*}
\]

\[3\]
Here, the sign $\rightarrow$ is the action on the inclusion of the document $D$ to the output $R$. The inclusion to the final output is carried out by a condition that can be either simple or compound, for example, "strong" and "fairly strong" linguistic proximity.

3. Conclusion

In the theory of computational semantic interpretation, the concept of semantic proximity based on a linguistic variable of fuzzy logic was introduced. The proposed transition from the values of the coefficients of semantic proximity of textual fragments to fuzzy variables. To calculate values of a linguistic variable, consistent decision rules were proposed and constructed, and a procedure for generating an open search in an information search was designed.

The proposed solutions will allow to organize an accurate and effective information retrieval process.

4. Acknowledgements

The work was carried out within the framework of the grant: 16-07-00241\18 of Russian Foundation for Basic Research.

References

[1] Feldman R and 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 and 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 and 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 and Shi H-B 2002 Improved feature selection approach TFIDF in text mining Machine Learning and Cybernetics, 2002. Proceedings. 2002 International. 2 944–946, IEEE.
[5] Hu K, Wu H, Qi K et Al. 2017 A domain keyword analysis approach extending Term Frequency-KeyWord 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 and Loukachevitch N V 2011 Multiple Evidence for Term Extraction in Broad Domains. RANLP 710–715.
[8] Efimenko I V, 2006, Obrabotka yestestvennoyazykovykh tekstov: ontologichnost’ v lingvistike i diskursivnost’ v izvlechenii znaniy (Processing of texts: ontology in linguistics and discourse in the extraction of knowledge) URL: http://www. raai. org/ressurs/papers/kiu-2006/doklad/ Efimenko_2.doc.
[9] Vishnyakov R and 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 and 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 R Y and Vishnyakov Y M 2013 Interpretatsionnaya model’ smysla tekstovogo fragmenta (Interpretation model of text fragment meaning) Tekhnologicheskiy institute Federal’noy gosudarstvennogo obrazovatel’nogo uchrezhdeniya vysshego professional’nogo obrazovaniya Yuzhnyy federal’nyy universitet v g. Taganroge 7(144) 2013
[12] Vishnyakov Y M and Vishnyakov R Y 2016 Vychislitel’naya semanticheskaya interpretatsiya tekstov nauchnotekhnicheskogo stilya (Computational semantic interpretation of texts of scientific and technical style) Sovremennyye naukoyemkiye tehnologii, 236–242, Obshchestvo s ogrаниченной отвественностью изданья Yuzhnogo federal’nogo universiteta v g. Rostov’pe” Izdatel’skiy Dom” Akademiya Yestestvoznaniya”.
[13] Vishnyakov Y M and Vishnyakov R Y 2017 Vychislitel’naya teoriya semanticheskoy interpretatsii tekstitv nauchno-tekhnicheskogo stilya yestestvennogo yazyka Izdatel’stvo “Nauchno-isledovatelskiye publikatsii”, 2017 219–226.