Identification of semantic objects in information stream

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 formulates and describes a formal approach to the identification of semantic objects in text streams. The reasoning is based on the theoretical provisions of the computational theory of semantic interpretation in terms of the computational representation of the semantics of text fragments and their comparison for semantic proximity. The proposed approach makes it possible to reveal the semantic traces of objects in text streams and to carry out their identification using them. A characteristic feature of the proposed approach is its use for identification of the grammatical structure of texts. We provide illustration for the given examples.

1. Introduction
The development of computer technology made it possible to process large amounts of data in real time, which in turn rose the emergence of systems to process audio, video and textual information. Systems began to implement the search, recognition and extraction of information by characteristic features that were set explicitly or implicitly in queries, e.g., search instructions, descriptions, etc. The processed data was often called a stream, since it was spread-out in time, and the matters of search (recognition, identification) were called objects... So, the terms video stream, video object, video recognition, etc. have moved from video surveillance systems to scientific vocabulary. We also expand these concepts to texts. Subsequently, an example of a text stream is a media news feed or a social network chat. A search object in a text stream can be, for example, a positive assessment of some movies. Since the stream is a text in a natural language, the search description of an object should also be linguistic in the form of words, phrases, text fragments, etc., setting in the aggregate the integral semantic content of the search object. The information about the search object contained in the text stream is linguistic and represents words, phrases, text fragments that form the semantic image of the object in the stream. Therefore, the presence of an object in the text stream can be judged by the semantic similarity of the search description of the object and its image in the stream. In this interpretation, the main question to be answered — what is the measure of the semantic similarity of the description of an object and its image in the stream, and in what way to implement their semantic comparison.

2. Definitions
To answer the the main question we start with a definition of an object and a linguistic description.
Definition 1: An object and its linguistic description, reflecting the internal semantics of the object, will be called a semantic object.

Definition 2: A linguistic description of a semantic object is the set $Q$ of its linguistic characteristics $q_i$:

$$Q = \{q_1, q_2, \ldots, q_n\}$$ (1)

The characteristic is given by a word or a chain of words built according to grammatical rules of the language. The chain of words, by definition, has an internal semantic integrity sufficient for search purposes. Linguistic descriptions for different purposes can differ significantly, therefore, their construction is not considered in this article. For each characteristic $q_i$, we introduce the measure of its presence $w_i$ in the text stream. Presence measure values are specified in the interval:

$$0 \leq w_i \leq 1$$ (2)

Definition 3: A trace of the characteristic $q_i$ in the text stream is the pair $(w_i, q_i)$.

Definition 4: The set $\{(w_i, q_i), i = 1 \ldots n\}$, representing in total an image of the semantic object in a text stream, is defined as the trace of the semantic object in the text stream. For a semantic object, we construct a recognition function $P$, for which we give each semantic characteristic a weight $p_i$ in such a way that the following relations hold:

$$\sum_{i=1}^{n} p_i = 1$$
$$0 \leq p_i \leq 1$$ (3)

We represent the recognition function itself as:

$$P = \sum_{i=1}^{n} p_i w_i$$ (4)

$$0 \leq P \leq 1$$

If $P = 1$, then a semantic object in a text stream is fully identified by all features, if $P = 0$, then a semantic object is not identified by any feature. In a similar way, it is possible to design a multiplicative recognition function, which is not fundamental. We should notice that the recognition function at the same time represents a measure of similarity between the linguistic description of a semantic object and its trace in a text stream. From the equation is obvious that the assignment of values to the weights $q_i$ is performed by the user when compiling a linguistic description of a semantic object and proceeds from his own internal considerations. In this case, more attention is paid to some characteristics, and less attention to others.

Definition 5: A semantic recognizer is a hypothetical system (device, program) that calculates the value of the recognition function in a text stream. Thereby it identifies the presence of a semantic object in the text stream from the linguistic description of a semantic object and its semantic trace.

3. Methods
We explain the introduced concepts with an example "international recognition of educational programs of Russian universities" in Russian language. Let a semantic object be represented by only one set of features: $q = $ международное признание образовательных программ российских вузов
The task is to find the trace \((w, q)\) of the given characteristic in the text stream, i.e., to determine to what extent there is a discussion of the educational programs of Russian universities in the context of the international recognition in the text stream.

In our case, the recognition function 2 looks like this:

\[
P = w \tag{5}
\]

Here, the method of forming and calculating the measure of the presence \(w_i\) of a feature \(q_i\) in the trace of the semantic object is crucial. This procedure involves comparing a feature \(q_i\) for semantic similarity of the text with other texts of its supposed patterns in a text stream. It essentially depends on an adopted model of semantics. Further we will discuss the current approaches of semantics modeling.

Frequency model. The most widely used in practice (information retrieval, classification, clustering, etc.). In the frequency model, the frequency properties of words in document texts are taken as a measure of relevance (semantic similarity, semantic proximity) [1]. However, taking into account the avalanche-like growth of information volumes and the increasing requirements of users to the accuracy and speed of information processing, a significant increase in the quality of processing within the frequency paradigm of relevance, apparently, can no longer be expected.

Nature-like technologies, collectively called artificial intelligence. Nowadays we can observe a renaissance of neural network technologies, the interest in which has revived due to the dramatically increased computing capabilities of computer technology [2, 3, 5]. Setting up a neural network for natural language processing comes down mainly to training phase. This phase provides a mapping on the structure and state of a network of a large number of words and their statistically stable connections [6]. These connections describe the processed texts. Therefore, high-quality training requires carefully formed representative training samples. It should be noted that the compilation of training samples, the assessment of their representativeness and the minimization of training time make an essential part of research on neural networks. A subjective expert opinion plays an important role here. Probably for this reason, to smooth out subjectivity in teaching a wide audience of researchers, projects from leading world corporations. Typically such projects are pre-trained. A good example of such a project is a Google’s BERT project [4].

With all the advantages of neural network technologies, there is a primary "negative" element, which is a learning phase. As noted above, neural network "understanding" is based on stable frequency characteristics of words and their connections, and they are formed in the learning process. In this case, training samples are deliberately assumed to be representative of the processed texts. However, the semantic content of the processed texts is not unchanged, it can change over time, the texts can be filled with new meanings. If these changes are slow enough, then the network can be trained regularly. With rapid changes in the semantic content of the processed texts, a complete retraining and reorganization of the network may be required. Therefore, training is an integral part of neural network technologies, but requires additional time and resources.

Ontologies. The semantics model in the form of ontologies has not found wide application in streaming. The main reason is the complexity of compilation, customization and reconfiguration [7].

This article proposes a different approach to model semantics. This approach was developed by the authors under the name "Computational theory of semantic interpretation" [10–13]. Within the framework of this approach was described the concept of a functional, expressing the meaning of a text and a procedure for calculating it. Comparison of text fragments is carried out indirectly by comparing the computational procedures of their functionals.

We consider the main provisions of the computational theory of semantic interpretation in the context of the goals and objectives of detecting semantic objects in text streams.
Let \( q \) be a string of words \( a_i \):

\[
q = a_1 a_2 \ldots a_n
\]  

(6)

This chain is an integral test fragment (a sentence or part of it). Integrity means that all words in the string \( q \) are included in phrases, forming a direct subordination relationship on the set of words of a text fragment. For simplicity, we will assume that this relationship is unambiguous.

We denote by \( S(a_i) \) the set of semantic meanings of the word \( a_i \) from 6, then the set \( S(q) \) of semantic values \( q \) in general form can be represented by some functional \( \Phi \), which is also a functional of a semantic:

\[
S(q) = \Phi(S(a_1), S(a_2), \ldots, S(a_n)).
\] 

(7)

Moreover, \( S(q) \subset S(a_i) \) if \( a_i \) is the main word of the chain \( q \), that is, the sense of the string \( q \) is a subset of the set of meanings of its main word \( a_i \).

Let’s select in \( q \) any phrase \( ab \), in which \( a \) is the main word, and \( b \) is the dependent one. Let us express this dependence by a formula:

\[
a : b,
\] 

(8)

In (8) the arrow shows the direction of the dependency. If the word \( a \) is the main word in several phrases with dependent words \( b_1, b_2, \ldots, b_p \), then this case is represented by the following formula, which is called the contextual connective of the word \( a \):

\[
a : \{b_1, b_2, \ldots, b_p\}.
\] 

(9)

The semantic meaning of a phrase is a subset of the semantic meanings of its main word, in the context of the meaning of the dependent word. Therefore, the semantic meaning of the phrase is represented by the following formula, similar to the conditional probability:

\[
S(a : b) = S(a)|_{S(b)}
\] 

(10)

For 10 the following relations are valid:

\[
S(a : b) \subset S(a); S(a : b) \neq S(b : a); S(a : a) = S(a)
\] 

(11)

We define the operation of contextual refinement of the meaning of the main word by dependent one as follows:

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

(12)

Here \( \overrightarrow{\cap} \) — 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 (12) to the context link (9), we obtain the following:

\[
S(a : (b_1, \ldots, b_p)) = S(a : b_1) \overrightarrow{\cap} S(a : b_2) \overrightarrow{\cap} \ldots \overrightarrow{\cap} S(a : b_p),
\] 

(13)

or otherwise:

\[
S \left( a : (b_1, \ldots, b_p) \right) = \cap_{i=1}^{p} S \left( a : b_i \right).
\] 

(14)

Here, the sign \( \cap \) — the operation of intersection.
By applying (8) — (14) it is possible to expand an expression of functional. To do so, we encode the words in the feature set $q$ (translation from Russian "international recognition of educational programs of Russian universities") in Latin letters as follows:

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

Taking into account the coding of the functional for $S(q)$, it is represented by the following expression:

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

Let’s discuss the comparison of tests for semantic similarity. The approach proposed in the computational theory of semantic interpretation is based on the comparison of computational procedures. This does not require specifying semantic meanings of words [11]. The computational procedures themselves are reduced to a certain comparison base, which is the notation for writing mathematical expressions in the form of reverse Polish notation (RPN) [14]. The notation does not contain parenthesis and is evaluated sequentially in one pass from left to right. In the works of the authors, the procedures for the transition to the representation of the functional of meaning in the RPN are constructed. So, in the RPN, the operation of contextual clarification of the meaning (12) takes the form as following:

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

Here, to maintain the dependence of words, the arrow above the operation changes direction. Also, the RPN of the contextual connection (13) with consideration of (16) takes the following form:

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

Here $p$ corresponds to the arity of the intersection operation.

Taking into account (16) and (17), the RPN of the meaning functional for $q$ (15) takes the form:

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

The steps of the RPN calculations (18) are as follows:

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

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

(iii) $S(q) = r_1 r_2 S(f) S(e) \cap S(d) \cap S(b) \cap \cap;$
(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) \( S(q) = r_1 r_6 \cap; \)

(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.

Thereafter, we consider a graphical representation of the computational procedure. Thus, we assign to each operation a graphical 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 (18) 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) (18) is presented in figure 1:

![Semantic schema of a textual fragment](image)

**Figure 1.** Semantic schema of a textual fragment \( q \)

4. Results and Conclusion
In this section we look at the comparison of strings for semantic similarity. Let it be required to compare semantically the set \( q \) with some set \( t \) of the form:

\[
t = b_1 b_2 ... b_m, \tag{19}
\]

Next, we introduce the criterion of semantic proximity (CSP) [5] as follows:
Here, \( C_{\text{prox}} \) 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 \( C_{\text{prox}} \). If \( C_{\text{prox}} = 0 \), then \( q \) and \( t \) are semantically not close. If \( C_{\text{prox}} = 1 \) there is a complete semantic match.

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

\[
C_{\text{prox}}(S(q), S(t)) \neq C_{\text{prox}}(S(t), S(q)).
\]

Let us represent the CSP in the form of the proportion of coinciding sense elements in the semantic scheme of the sets \( q \) and \( t \) to the total number of sense elements \( q \):

\[
C_{\text{prox}}(S(q), S(t)) = \frac{p}{m}.
\]

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.

Consider the process of identifying a semantic object by its characteristic \( q \) in a certain text stream. The text stream contains sets of the form:

\( t_1 = \text{Правительство России проводит политику, направленную на международное признание образовательных программ российских вузов} \)

The sentence \( t_1 \) in the example can be translated into English as "The Russian government pursues a policy aimed at the international recognition of educational programs of Russian universities".

\( t_2 = \text{образовательные программы российских вузов нуждаются в международном признании} \)

The sentence \( t_2 \) in the example can be translated into English as "Educational programs of Russian universities need international recognition".

\( t_3 = \text{международное признание образовательных программ поднимает репутационный рейтинг российских вузов} \)

The sentence \( t_3 \) in the example can be translated into English as "International recognition of educational programs raises the reputation of Russian universities".

\( t_4 = \text{международное признание вузов увеличивает возможности России по привлечению иностранных студентов в вузах} \)

The sentence \( t_4 \) in the example can be translated into English as "International recognition of universities enhances Russia's ability to attract foreign students".

\( t_5 = \text{международные программы академических обменов и участие в них российских ученых способствуют улучшению образовательного процесса вузов России} \)

The sentence \( t_5 \) in the example can be translated into English as "International programs of academic exchanges and participation of Russian scientists in them contribute to improving the educational process of Russian universities".

Let’s check the feature \( q \) for semantic closeness to the strings \( t_1, t_2, ..., t_5 \) or, in other words, determine whether these strings represent a trace of a semantic object in a text stream.

1. Comparison for semantic proximity of the string \( q \) with the string \( t_1 \).

\[ \text{Правительство России проводит политику, направленную на международное признание образовательных программ российских вузов.} \]

Here, the string \( q \) is a sub string in the string \( t_1 \) (in bold) and its semantic scheme completely fits into the semantic scheme of the string \( t_1 \), therefore \( C_{\text{prox}}(S(q), S(t_1)) = 1 \).

2. Comparison for semantic proximity of the string \( q \) with the string \( t_2 \).

\[ \text{образовательные программы российских вузов нуждаются в международном признании} \]
In $t_2$, sub strings that match the fragments $q$ are also in bold. In the semantic diagram $q$ shown in 2, matching elements of meaning are highlighted in bold lines.

![Figure 2](image-url)

**Figure 2.** The comparison result of the proximity of the $q$ and the $t_2$.

The semantic scheme of the sample contains 5 coinciding elements of meaning, and therefore its value $C_{prox}(S(q), S(t_2)) = 5/7$ is still quite large.

3. Comparison for semantic proximity of the string $q$ with the string $t_3$, see 3.

Международное признание образовательных программ поднимает репутационный рейтинг российских вузов

In this example is obvious that the semantic schemes $q$ and the $t_3$ have only three coincident elements of meaning, therefore $C_{prox}(S(q), S(t_3)) = 3/7$.

4. Comparison for semantic proximity of the string $q$ with the string $t_4$, see 4.

Международное признание вузов увеличивает возможности России по привлечению иностранных студентов.

![Figure 3](image-url)

**Figure 3.** The comparison result of the proximity of the $q$ and the $t_3$. 
Figure 4. The comparison result of the proximity of the $q$ and the $t_4$.

Figure 5. The comparison result of the proximity of the $q$ and the $t_5$.

Figure 4 shows that the semantic diagrams of the $q$ and $t_4$ have only one matching element and $C_{prox}(S(q), S(t_4)) = 1/7$.

5. Comparison for semantic proximity of the string $q$ with the string $t_5$, see 5.

Международные программы академических обменов и участие в них российских ученых способствуют улучшению образовательного процесса вузов России.

The result of comparison of the $q$ and the text of $t_5$ is represented in figure 5.

The semantic schema of the $q$ does not have the same elements of the meaning as in the text of the $t_5$, therefore $C_{prox}(S(q), S(t_5)) = 0$.

And so, when analyzing the text stream in relation to the linguistic features $q$, four text fragments $t_1$, $t_2$, $t_3$, $t_4$ with a nonzero semantic proximity of 1, 0.71 were identified; 0.43; 0.14 respectively.

Now let’s construct the trace of a semantic object in the form of a set $\{(w_i, q_i), i = 1 \ldots n\}$. For this, it is necessary to form measures of the presence of $w_i$ for each of the linguistic feature.

While analyzing a text stream for a linguistic features $q_i$, a set of its nonzero semantic affinities
is constructed:

\[ \{C_{\text{prox}}(S(q_i), S(t_k)), k = l_1, l_2, l_1 ... l_r \} \] (23)

Then, for the degree of presence \( w_i \), you can choose the maximum value of the criterion for the proximity of this feature:

\[ w_i = \text{MAX}\{C_{\text{prox}}(S(q_i), S(t_k)), k = l_1, l_2, l_1 ... l_r \} \] (24)

Or the average value of the proximity criterion:

\[ w_i = \text{\{C}_{\text{prox}}(S(q_i), S(t_k)), k = l_1, l_2, l_1 ... l_r \} \] (25)

Further, the formation of the recognition function is carried out on the basis of (2). For the semantic object considered above, the presence measure \( w \) takes the value 1 in the case of its determination according to (24) as the maximum value, \( w = 0.57 \) - in the case of the average value according to (25). Accordingly, the recognition function will have values of 1 or 0.57, respectively. In any case, based on these results, it is possible to draw a conclusion about the degree of presence of the semantic object in the text stream.

We formulated a formalized approach to identify semantic objects in text streams. The approach was based on the provisions of the computational theory of semantic interpretation in terms of representing semantics and comparing text fragments for semantic proximity. The proposed solutions make it possible to reveal the semantic traces of objects and, by them, identify them in text streams, which makes it possible to build effective semantic recognizers.

References

[1] 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 International,2:944–946, IEEE
[2] Feldman R, Sanger J, 2007, The text mining handbook: advanced approaches in analyzing unstructured data Cambridge university press
[3] 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
[4] Devlin J, Chang MW, Lee K, Toutanova K, 2018 Bert: Pre-training of deep bidirectional transformers for language understanding arXiv preprint arXiv:1810.04805
[5] Tseng Y-H, Lin C-J, Lin Y-I, 2007, Text mining techniques for patent analysis Information Processing & Management, 43(5):1216–1247, Elsveir
[6] 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
[7] Cruse A, 2011, Meaning in language: An introduction to semantics and pragmatics Oxford University Press UK
[8] Dobrov B V, Loukachevitch N V, 2011, Multiple Evidence for Term Extraction in Broad Domains. RANLP, 710–715
[9] Efimenko I V, 2006, Obrabotka yestestvennoyazykovykh tekstov: ontologichnost’ v lingvistike i diskursivnost’ v izvlechenii znanii (Processing of texts: ontology in linguistics and discourse in the extraction of knowledge) URL: http://www. raai. org/resurs/papers/kii-2006/doklad/Efimenko_2.doc
[10] 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
[11] 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
[12] Vishnyakov R Y, Vishnyakov Y M, 2013, Interpretatsionnaya model’ smysla tekstovogo fragmenta (Interpretation model of text fragment meaning) Tekhnologicheskiy institut Federal’nogo gosudarstvennogo obrazovatel’nogo uchrezhdeniya vysshego professional’nogo obrazovaniya Yuzhnnyy federal’nyy universitet v g. Taganroge,7 (144):2013
[13] Vishnyakov Y M, Vishnyakov R Y 2016, Vychislitel'naya semanticeskaya interpretatsiya tekstov nauchno-tekhnicheskogo stila (Computational semantic interpretation of texts of scientific and technical style) Sovremennye naukoyemkiye tehnologii, 236–242. Obshchestvo s ogranichennoy otvetstvennost'yu Izdatel'skiy Dom" Akademiya Yestestvoznaniya"

[14] 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