Analysis of associations in texts for determination of associants' relation types

Alexander A. Kharlamov[1], *Rodmonga K. Potapova*[2], *Tatyana V. Yermolenko*[3]

[1)] Institute of Higher Nervous Activity and Neurophysiology of the Russian Academy of Sciences, Moscow
kharlamov@analyst.ru

[2)] Moscow State Linguistic University, Moscow
rkpotapova@yandex.ru

[3)] Institute of Artificial Intelligence Problems, Donetsk, Ukraine
etv@iai.dn.ua

Abstract

The homogeneous semantic (associative) network built using the neural network technology TextAnalyst consists of pairs of associations as they occur in the text. The analysis and classification of such associations obtained from specific texts show that in fact these associations are classified into classes of predicates' actants. The resulting conclusion suggests that using comprehensive linguistic analysis for construction of extended predicate structures for simple fragments of sentences, it is possible to obtain information on relations between key concepts of a homogeneous semantic (associative) networks, that is, to move from homogeneous (associative) semantic networks to heterogeneous ones. Moreover, these relations between key concepts in this case can be marked using automatic analysis. The replacement of associative relations in a semantic network by diverse ones, in addition to improving the quality of classification and text abstracting, will allow for automatic construction of non-homogeneous (heterogeneous) semantic networks that so far has been done manually by experts.

1. Introduction

TextAnalyst, a neural network technology for constructing the semantic "map" of a text (a homogeneous semantic (associative) network), which presents formation of the frequency "map" of a text (a set of word pairs as they occur in sentences (or other semantic fragments) followed by renormalization of the frequency of occurrence for key concepts (words and collocations) of a text into their semantic weights, allows for such useful functionality of the text analysis as referencing and automatic comparison (and hence classification) of texts [1]. In neurophysiology [2, 3] and neuroinformatics [1] one can find some descriptions of the human brain mechanisms responsible for step-by-step processing sensory data (including text data). This information processing in the human brain is taken as the basis for the construction of mechanisms of automatic semantic analysis of texts in the TextAnalyst technology.

During this processing, first in the cerebral cortex of the brain, dictionaries of text units are revealed at various levels, including root stems of full words at the lexical level, and pairs of words compatible within their meanings - at the semantic level, which (the text units) are provided with information about the frequency of their occurrence in the text [1]. Then, in the hippocampus, these text units are combined into networks characterizing their co-occurrence within some situations (texts) [3]. And the frequency of occurrence for these text units within the whole text is renormalized into their semantic weight according to their significance in the situation [1].

The semantic "map" of a text (associative network) is generated automatically by using the TextAnalyst technology in three stages. At the first stage a preprocessing of a text is implemented with removal of stop words, functional and commonly used words as items that do not have any semantic meaning. Simultaneously, the morphological analysis of the remaining words (or rather their stemming) is performed in order to make the analysis more robust, and the semantic network being formed more compact (and therefore more convenient to visualize) by combining vertices of various word forms of one word.

At the second stage the frequency "map" of the text is formed. The frequency of occurrence is calculated for root stems (obtained as the stemming results), as well as the frequency of pairwise occurrence of the root stems in sentences. These pairs of words can be combined into a network (in this case, an associative one, since the relations between the concepts in a pair are associative, that is, taking into account only their co-occurrence in the sentences of the text).

This combining of word pairs into a network allows the third stage: an iterative procedure for renormalization of the key concepts' frequency of occurrence into their semantic weight. In this case, the concepts of the network associated with a larger number of other concepts with larger weights, increase their weight. Other concepts lose it proportionally. As a result of the renormalization larger weights belong to the concepts, the frequency of occurrence of which can be low, but which provide the main content (general sense) of the text.

A homogeneous semantic network is convenient for the purpose of automatically extracting the main content (sense) of the text (abstracting), as well as for comparison (and therefore, classification) of texts [1]. However, a heterogeneous semantic network (the relations in which are marked by relation types) contains more information, and therefore, is more convenient for the analysis of texts: it allows to more accurately identify the main content (sense) of the text and to compare texts more accurately. Moreover, it provides the possibility to implement a pragmatic analysis of texts [4]. But this requires comprehensive linguistic analysis of individual sentences of the text in order to identify the extended predicate structure of simple sentences of text fragments that determines values of relations of the text key concepts [5].

To confirm the correctness of the change from a homogeneous semantic network to a heterogeneous one, it was necessary to conduct a series of experiments which showed that associations revealed in the text by using the TextAnalyst technology, can indeed be replaced by specific relations between the concepts, since these associations are classified into groups corresponding, on the one hand, to valences of verbs, and on the other hand – to a set of predicate relations. This work was carried out on a set of news texts (of course, sources of texts can be different) of small volume (to reduce manual handling).

2. Analysis of associations in texts for determination of associants' relation types

The term "association" in the context of this paper refers to co-occurrence of a word pair in a semantic fragment of a text, for example, in a sentence. Since the paper considers the automatic analysis of texts, to identify such associations the TextAnalyst technology for automatic semantic analysis of texts was used. When analyzing a text an associative network is formed that represents a combined set of word pairs extracted from the analysis of the text. The second word of the first pair is combined with the first word of one of the available pairs; then the second word of this second attached pair is attached to another pair, in which the first word is identical to the second word of the previous pair, etc. These word pairs are identified at the stage of constructing a frequency "map" of the text and represent all pairwise occurrences of words in the text sentences: the first word with the second one; the first word with the third one; the first word to the fourth one, etc.; the second word with the third one, the second word with the fourth one, etc.; the third word with the fourth one, etc.

The material thus prepared was further processed manually in the experiment with the use of comprehensive linguistic analysis in order to mark relations in word pairs and then classify these pairs by relation types.

2.1. Comprehensive linguistic analysis

The mechanism of comprehensive linguistic analysis includes steps of morphological analysis and syntactic and semantic analysis of a separate sentence [4]. The morphological analysis of text implies the use of a combined dictionary and non-dictionary approach [5]. Morphological information obtained during the morphological analysis is used in the syntactic and semantic analysis for fragmentation of sentences into simple fragments to remove syntactic homonymy in the analysis of these fragments, and to form some templates of minimal structural patterns of sentences that describe the predicative minimum of this sentence. To identify extended predicate structures of simple fragments of text sentences, the dictionary of verbs' valences is used [6].

2.1.1. Morphological analysis

Since most of the words of a text is an unchanged basis of language and is covered by vocabulary within a hundred thousand words, and the other, more rare, but no less important part of the lexicon is constantly updated and has no clearly defined boundaries in principle, especially with regard to proper names and word-formation variants of known words, the morphological analysis includes methods with both declarative and procedural objectives.

The declarative morphological analyzer uses the full dictionary of all possible word forms for each word. In addition, each word form is provided with complete and unambiguous morphological information, which includes both fixed and variable morphological parameters. The objective of the morphological analysis is reduced to finding the right word form in the dictionary. If the word is not found, then procedural methods are used, where the stem and the affix of each word are identified, and the dictionary contains only stems of words along with relations to the corresponding rows in the dictionary of affixes [7, 8].

2.1.2. Analysis of the syntactic and semantic level

The semantic and syntactic analysis of a sentence is performed in several steps: fragmentation of a sentence; combining homogeneous fragments; hierarchization between fragments of different types; combining fragments into simple sentences; building of simple syntactic groups within the fragments; identification of the predicative minimum of each simple sentence; identification of the remaining members of a simple sentence that are actants of the identified predicate; construction of syntactic groups, in which the predicate actant is the main word.

Syntactic rules define relations between words (segments) in the predicative form. Depending on the type of segments and type of the subordinative conjunction, and using heuristic rules it is possible to implement multiple operations: subordinance, homogeneity, implication, and conjunction. The result is decomposition of complex sentences into simple sentences connected with coordinating or subordinating conjunctions.

Next step is construction of simple syntax groups within each simple sentence and identification of the predicate nucleus. Simple syntax groups include groups at the attribute level, groups with prepositions and comparative constructions.

A set of simple sentences of the Russian language is given by a list of the minimum structural patterns of sentences describing the predicative minimum of the sentence.

In all segments of the sentence that are not nested and homogeneous, a serial search of a suitable template of the minimal structural pattern of the sentence is performed. In accordance with the pattern found, each main member of the sentence is assigned an appropriate value.

Then, the extended predicate structure of simple sentences is obtained, and predicate's valence nests are filled [9]. Identification of the remaining members of a simple sentence (the remaining semantically significant objects and attributes) is implemented by sequentially comparing the words of the sentence with the verb actant structure, which requires the use of the dictionary of verbs' valences.

2.2. Neural network modeling by the example of text information analysis

To identify associations, an array of texts from the news feed [Archive NEWSru.com] was used, the fragment of which is given below.

«Архив NEWSru.com:: 9 сентября 2013 года TXT PDA MOB Понедельник, 9 сентября 2013 г. 18+ Москва предложила Дамаску "химическое разоружение". Сирии предлагается передать имеющееся химоружие под международный контроль с последующей его ликвидацией. Дамаск поприветствовал "мудрость российского руководства", в Лондоне и ООН идею также поддержали, в США отнеслись скептически.

Навальный выступил перед тысячами сторонников, объявив о рождении в России политики 08:50

Путин в преддверии Олимпиады собрал Совбез по вопросу терроризма на Северном Кавказе 01:12. Мечеть "Сердце Чечни" досрочно признана символом страны в конкурсе "Россия-10" 13:55.

Последнее обновление: 09:45. ОНФ хвастается первыми выявленными нечестными госзакупками на миллиарды рублей. В рамках проекта "За честные закупки" ОНФ отменил через ФАС две нечестные госзакупки: одна на 3,9 млрд. рублей, вторая на 85 млн. рублей. На сайте, который начал работать 1 сентября, на рассмотрении 28 сомнительных закупок на общую сумму 19 567 543 929 рублей».

As a result of the processing the text array using the TextAnalyst software, a semantic network was obtained, a fragment of which is presented in Table 1.

Table 1: Fragment of a homogeneous semantic network represented by word pairs (superordinate-subordinate words)

№	Superordinate word	Frequency	Weight of subordinate word	Subordinate word
1	Президент	2	49	Израиль
2	Президент	2	33	лидер
3	Москва	2	31	бизнес
4	Москва	2	31	Киев
5	Москва	2	31	мигранты
	Москва	2	31	партия
	обвиняемый	3	51	суд
	обвиняемый	2	35	водитель
...				

Within this work we consider two types of syntactic relations - predicative and syntagmatic ones. In turn, these types of relations include several types of syntactic relations. Predicative relation types of relationships are "predicate-actant" relations, which correspond to the valence slots of a predicate where actants act as semantic cases: subject, object, addressee, tool and locative (initial, final, intermediate). Types of subordination are used as types of a syntagmatic relation, namely: attribute, genitive, comparative construction, etc.

An example of a fragment of a heterogeneous semantic network with marked types of syntactic relations is presented in Table 2.

Table 2: Fragment of a heterogeneous semantic network represented by word pairs (superordinate-subordinate words) and their syntactic relations

№	Superordinate word	syntactic relations	Subordinate word
1	Президент	predicative	Израиль
2	Президент	predicative	лидер
3	Москва	predicative	бизнес
4	Москва	predicative	Киев
5	Москва	predicative	мигранты
	Москва	predicative	партия
	обвиняемый	predicative	суд
	обвиняемый	syntagmatic	водитель
...			

As a result of using the mechanism of comprehensive linguistic processing, the following classes of relations were identified between the key concepts in the texts previously processed (see Table 3).

Table 3: Frequency of occurrence of syntactic relation types identified in texts from the news feed using comprehensive linguistic analysis

№	Type of syntactic relations	Frequency of occurrence,%	Context
1	subject-predicate (разочаровать)-object	18	… **президент** … разочаровал … Израиль
2	subject-predicate (пожелать)-object	18	… лидер пожелал **президенту** …
3	address predicate (предложить)-object	7	… предложит **Москве** бизнес …
4	subject-predicate (намекнуть)-addressee	9	**Москва** намекнула Киеву …
5	locative-predicate (ловить)-object	6	… **Москве** … ловят мигрантов…
6	addressee-predicate (предложить)-subject	9	… партия предложит **Москве** …
7	object-predicate (арестовать)-subject	18	Суд арестовал **...** **обвиняемого**…
8	attributive	10	Водитель …, **обвиняемый**…
…			

Note to the table 3: valence slots of superordinate words (bold type) are given in the first place.

3. Discussion

What is of interest concerning heterogeneous semantic relations between key concepts in sentences of a text that can be used to improve the automatic semantic processing of texts? Introduction of heterogeneous marking of relations in a semantic network instead of associations only leads, on the one hand, to the semantic network breakdown (and consequently to reduction of the analysis robustness), but, on the other hand – to a more accurate comparison of individual fragments of networks when comparing and classifying texts. In order to provide the possibility to use comprehensive linguistic analysis for formation of heterogeneous semantic networks, let us compare the structure of homogeneous and heterogeneous semantic networks.

The structure of a homogeneous semantic network is determined by pairwise associations of key concepts in a text. A set of such pairs exhaustively represent the content of the text - its contents is represented in the form of its semantic “map”. Let us structure the network more compactly: combine all pairs with the same superordinate word. We obtain the so-called “stars” – subordinate words depending on the superordinate words (the so-called associants) become their semantic features. Now the primary frequency network is constructed as a group of stars.

Such a representation clearly shows that associative and syntagmatic relations in a meaningful text are very highly correlated: the presence of associations is determined by their syntagmatics, and the syntagmatics is determined by semantic dependencies. And the classification obtained as a result of the experiment shows that association classes naturally break down to, on the one hand, to a set of predicate relations (each verb has its own set), and on the other hand – to a set of relations determined by valences of these same verbs.

In other words, one can imagine stars, which are formed in the process of analysis of simple fragments of text sentences, in which the superordinate word is the subject, the primary and secondary objects and their attributes are subordinate words of the star, and a predicate relation and verb valences describe relations, respectively, of the primary and secondary objects and their attributes with the subject.

The network formed of these stars becomes heterogeneous, as in addition to the key concepts it contains relations marked by types. This network is, however, primary as it is a frequency “map” of the text. After its renormalization and replacement of key concepts’ frequency of occurrence by their semantic weights, the network becomes a heterogeneous semantic network.

A heterogeneous semantic network for further analysis of texts (for example, to obtain some abstract, to compare (classify) texts) is a more delicate material than a homogeneous semantic network.

4. Conclusions

The results presented in this work show that associative pairs detected by automatic analysis of texts with the help of the TextAnalyst technology are classified in a variety of classes described by some types of relations, on the one hand – by predicate ones, on the other – by relations of verbs’ valences. Such clustering of the results suggests the possibility of automatic detection of relation types between concepts in a semantic network using methods of comprehensive linguistic analysis. That, in turn, provides the possibility to improve the quality of automatic analysis in the TextAnalyst technology regarding formation of text abstracts, as well as comparison and classification of texts.

5. Acknowledgements

The work was completed as part of the research “Study of the mechanism of associative relations in human’s verbal and cogitative activity using the neural network modeling method in the analysis of text information” (supported by the Russian Foundation for Basic Research, Grant 14-06-00363).

6. References

[1] Kharlamov A.A. *The neural network technology of information representation and processing (natural representation of knowledge).* Publ.House «Radiotekhnika», Moscow, 2006 (in Russian).

[2] Hubel D.H. *Eye, brain and vision.* Scientific American Library, New York, 1988.

[3] Vinogradova O.S. *Hippocampus and memory.* Publ.House «Nauka», Moscow, 1975 (in Russian).

[4] Alexander A. Kharlamov, Tatyana V. Yermolenko, Andrey A. Zhonin. *Text Understanding as Interpretation of Predicative Structure Strings of Main Text’s Sentences as Result of Pragmatic Analysis (Combination of Linguistic and Statistic Approaches) //* International Conference SPECOM 2013, Plzen, Czech Republic, September 2013.

[5] Potapova R.K. *Technologies of Natural Language Processing in Science and Industry.* Publ. House «INION RAS», Moscow, 1992 (in Russian).

[6] Bondarenko E.A., Kaplina O.A. *The principles of automated natural-language texts processing: the valence approach* // Artificial Intelligence. — 2013. — N1. — C. 80-90 (In Russian).

[7] Dorokhina G.V., Pavlyukova A.P. *The module of morphological analysis for Russian words* // Artificial Intelligence. – 2004. – № 3. – С. 636-642 (In Russian).

[8] Dorokhina G.V., Trunov V.Yu., Shilova E.V. *The module of morphological analysis without a Russian language dictionary* // Artificial Intelligence. – №2. – 2010. – С.32-36 (In Russian).

[9] Kharlamov A.A., Yermolenko T.V., Dorokhina G.V., Gnitko D.S. *Singling out principal parts of a sentence as predicative structures using minimal structural patterns* // Rechevye tekhnologii. — 2012. — №2. — С.75-84 (In Russian).