Organisation natural language sentences using logic-linguistic models

Nastya Vavilenkova

Department of Computer Management Systems,
National Aviation University, UKRAINE, Kiev,
Komarova street 1, E-mail: a_vavilenkova@mail.ru

Abstract – The article considers the approaches used for semantic analysis of text information: finite state transition network and content analysis. It is proposed to use logic-linguistic models for organisation natural language sentences.

Key words – natural language, logic-linguistic model, transition network, semantic modeling, predicate, semantic-syntactic relations, text information.

I. Introduction

The overall base of semantic analysis methods that can detect semantic relations between words is a thesaurus of language. At the mathematical level it is a directed graph, whose nodes are the words in their basic word forms and arcs define the relations between words and can also display a number of features. Thus thesaurus defines a set of binary relations on the set of words of natural language. Each sentence of natural language has structured minimum that can be represented as a logic-linguistic model of text information. This model is based on predicate logic. Predicate is in the predicative relation to the subject, it is able to acquire different modal values. Predicate is a meaningful aspect because there are not only formal types of predicate sentences, but also semantic types of predicate. Deep structure of any natural language sentence corresponds to its semantic interpretation. That is, the semantic component should contain rules that transform the underlying structure of sentences generated by the syntactic component in their semantic representation [1]. When person speaks, he understands the contents of any sentences from any set of sentences, performing transaction association meaning of the words in the content of phrases and sentences. The rules of semantic components have to execute this unique procedure: to build the content of complex aggregate from the contents of its components.

Semantic analysis involves procedures aimed at automatic semantic processing of text and creation on its basis new linguistic objects. Semantic analysis is an algorithm that allows us to represent the semantic (content) structure of sentence and text as a strict formal system through analytical exploration of the relations between individual objects and events from the subject area. Semantic component is a set of concepts represented in words and phrases that are related to each other in content. These concepts form a semantic dictionary where described units are grouped not formally (in alphabetical order), but according to the semantic sets (classes, groups, etc.). That dictionary is based on a hierarchical system of concepts representing its different semantic relations and it is necessary source of semantic information for applications of automatic text processing. The applying of such systems requires component that executes semantic analysis and work with the content of text. The purpose of semantic analysis is to determine the content characteristics for each word and phrase as a whole. Difficulties arise due to semantic ambiguity. Often, to remove this ambiguity, it is necessary to use "semantic articles" related to each other within the semantic network [2]. Analysis of the relations within the semantic network provides with an opportunity to get information that is obviously missed in phrase but without this information any adequate understanding of the phrase is impossible. Difficulties of such implementing are associated with a large amount of semantic networks and multiplicity of analysis. Representation of the sentence that obtained at the stage of semantic analysis is called semantic graph of sentence.

II. The method of augments transition networks

A finite state transition network is represented by set of nodes and directed arcs connecting them. These nodes correspond to nonterminal symbols and arcs to terminal symbols. Sentence is a minimal and basic communication unit of the language. Sentences should be holistic and transmit information across the complexity of dependencies and relations [3]. Syntactic relations in sentences are called according to the function of dependent member of sentence: identification relations (between the subject and the attributive, complement, adverbial), adverbial relations (between predicate and adverbials), complement relations (between predicate and complement), predicate relations (between subject and predicate). Based on this classification and the assumption that each sentence in natural language has a certain structural minimum, it is possible to build a finite state transition network for any sentence (Fig. 1) [4].

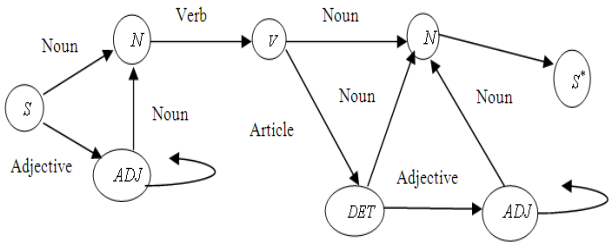


Fig. 1. Finite state transition network

Where S is initial node, and S^* is final node.

Terms of use of the finite state network:

1. We have to choose one of the directed arcs, which comes from this node and go through it.

2. When the arc is passed, we have to choose one of the terminal symbols subset corresponding to that arc.

3. Continue the process until a node S^* would be reached.

So, it is possible now to write an order of consideration of network nodes (Fig.1) for the sentence «Digital logic circuits require two levels of signal».

Procedure for consideration of nodes would be:

$$S \rightarrow ADJ \rightarrow ADJ \rightarrow N \rightarrow V \rightarrow ADJ \rightarrow N \rightarrow DET \rightarrow N \rightarrow S^*.$$

Thus, using a specific network (Fig. 1), we are able to reach final state S^*, so the sentence «Digital logic circuits require two levels of signal» is perceived. However, the proposed finite state transition network (FTN) is not universal. If we change the word order within the natural language sentence or increase the number of a certain type members of sentences, the network will not be able to bring the user to the final state S^*. For sentence to be perceived, cycles should be input into the FTN for almost every of its terminal symbols. Thus the network will have an infinite number of states.

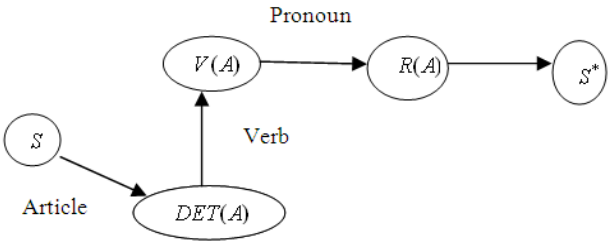
In addition, the proposed network is able to work with rules of recursive origination. Within these rules, a single left symbol reconstitutes right symbol, for example $Cq \rightarrow A_1qA_2$. Since, there is no way to implement recursion as a part of FTN, a clause of natural language can be described by a single network.

That is, the recursion can be done by extending the model of finite network and enabling a FTN to call the second such network. This procedure should be able to run in any node. In this case, the grammar is represented by set of FTNs, each of which corresponds to a grammatical analysis of natural language sentences. Then the analysis of T chain is the following algorithm:

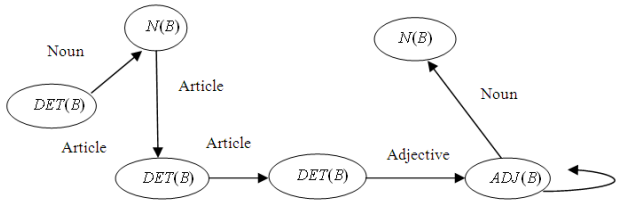
1. We start examination of terminal symbol S in the initial network.
2. We pass an arc and perceive terminal symbol at each step.
3. If necessary, some nodes, instead of passing the original arc should assigns control to another FTN, referring to the grammar.
4. A called network starts analysis with its own original symbol $S^{'}$, using the next symbol of T chain as its first input symbol.
5. In turn, if necessary, called network can initiate related FTN, etc.
6. When reaching the final node $S^{'*}$, called network assigns control to the initial network, which has been analyzing a input chain T.
7. The process continues until you reach the final terminal symbol of original network S^*.

While studying of how the transition network analyzes the natural language sentences, it is necessary to distinguish between two components of the algorithm: the actual network and program management. Management program is responsible for memorizing word which is read from the input chain T and also for a sequence of calls and network's place active at the moment. In FTN without recursion, controlling program simply checks whether the read word is label for one of the arcs, which comes from the currently active node (terminal symbol).

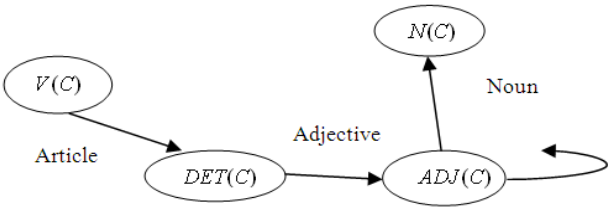
Consider the example of a set of networks for FTN, which has the ability to call other networks like procedure. Let's analyze the sentence: «The micro-processor is an integrated circuit which has the properties of a complete central processing». For the analysis of a input sentence it is proposed to use a set of networks M={(A),(B),(C),(D)} (Fig.2).



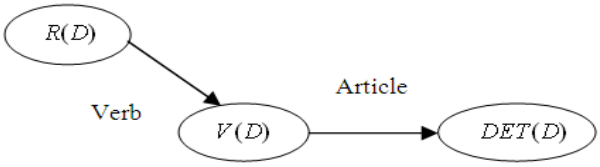
a) Network A



b) Network B



c) Network C



d) Network D

Fig.2. Set of FTNs

1. The analysis begins with the state S of the network (A).
2. The control is passed to DET (A) and the input word will be «The».
3. The network (B) contains a set of possible states for describing the noun, which is the next word in the chain. However, called network (B) for word «microprocessor» has only one proper arc *DET (B) → N(B)*, where DET(B) – initial terminal symbol of network (B).
4. Procedure of consideration of the chain at the moment is *S→DET (A)→N (B)*, current word «is».
5. From the initial FTN's node V(A) the network (C) is called whose initial node is V(C).
6. The transition on network (C) occurs, and result will be as the following sequence of stages: *V(C)→DET(C)→ADJ(C)→N(C)*.
7. After returning to the original network (A), the active node R(A) will be the word «which», and the sequence of stages that led to this node:

S→DET(A)→N(B)→V(C)→DET(C)→ADJ(C)→N(C).

8. From node R(A) the network (D) is called, and initial terminal symbol is R(D).
9. Analysis of a set of states of network (D) makes possible to define the following sequence of stages:

R(D)→V(D)→DET(D) then comes another call of network where the initial state DET(B), i.e., the word «the».

10. After analyzing the network (B) a sequence of stages is formed as follows:

$$DET(B)\rightarrow N(B)\rightarrow DET(B)\rightarrow DET(B)\rightarrow ADJ(B)\rightarrow \rightarrow ADJ(B)\rightarrow N(B).$$

11. From the network (B) it is returned first to the called network (D), and then to original network (A), which following state is finite S^*.

12. Thus, the call of FTN within input set led to the following sequence of stages:

$$S\rightarrow DET(A)\rightarrow N(B)\rightarrow V(C)\rightarrow DET(C)\rightarrow ADJ(C)\rightarrow \rightarrow N(C)\rightarrow R(D)\rightarrow V(D)\rightarrow \rightarrow DET(B)\rightarrow N(B)\rightarrow DET(B)\rightarrow DET(B)\rightarrow ADJ(B)\rightarrow \rightarrow ADJ(B)\rightarrow N(B)\rightarrow S^*.$$

This approach is generally correct, but the implementation is very labor-intensive, requires large amount of actions and takes much time. The larger is the network, the harder is to use it. The implementation approach using FTN is possible, if the correct number of nodes in the network is provided. This option is available when working with a specific type of sentences. In his work on transformational grammar, Chomsky [5] notes that in formal linguistics we research the concept, i.e. order the structure of natural language sentence, rather than its execution. Thus, the practice of using the proposed approach shows the impossibility of using it for semantic structuring of sentences. And as the structure of natural language sentences is very diverse, the approach using FTN is not rational and universal.

III. Organization natural language sentences using logic-linguistic models

The considered method of augmented FTN gives a possibility to trace visually the syntactic and semantic relations in complex natural language sentences. By a similar principle we can identify the relation between sentences in any text. Nowadays, to research texts a content analysis is used. Its main purpose is to identify the content of text arrays to further meaningful interpretation of the discovered numerical patterns [6]. The basic idea of content analysis is to discover the procedures by which we can find corresponding indicators studied phenomena and characteristics in the text. Content analysis is used as the primary method aims to obtain the most important information about the subject area. This method, used in combination with others, as an auxiliary procedure for processing data obtained in other researches.

The object of content analysis is the content of various electronic documents interpreted through statistical calculation of meaningful units: concepts expressed in words and terms, themes, expressed in the form of paragraphs of texts and articles [7].

A disadvantage of content analysis is that the researcher should take into account not only mentions that may encounter in the text, but also elements of its contextual use. For this purpose a detailed system of rules for each case use should be developed. Also, the positive and negative ratings are assigned to key content units manually and not automatically; content units are also discovered not automatically but by experts. Searching of natural language sentences by building its logic-linguistic models allow us to analyze the content of sentences: by detecting similar elements and analyzing predicate variables of constructed models.

Example usage the scheme of knowledge extraction from sentence by means logic-linguistic model: «Discipline studying the models and methods of knowledge extraction».

1) After receiving characteristics of each word and after using the rules of production models it is possible to define functional relationships between words, ie:

- Studying the models;
- Studying methods;
- Models of knowledge;
- Methods of knowledge;
- Knowledge extraction.

2) Identify syntactic roles that the words performs in sentence:

«discipline» – subject x_1 – predicate variable subject;

«studying» – predicate P – predicate;

«models» – object x_2 – predicate variable argument;

«methods» – object x_3 – predicate variable argument;

«knowledge» – object x_4 – predicate variable argument;

«extraction» – object x_5 – predicate variable argument.

3) The logic-linguistic model of natural language sentences is forming as follows:

$$P\left(x_1, x_2[x_4[x_5]] \,\&\, x_3[x_4[x_5]]\right),$$

$$studying\begin{pmatrix} discipline, \mathrm{mod}\, els[knowledge[extraction]] \,\& \\ \&methods[knowledge[extraction]] \end{pmatrix}.$$

For example, we have a set of sentences: «Robots can also tell the difference between two temperatures. Ukraine is a sovereign state with its territory, high and local bodies of state power, government. He was one of the greatest scientists and thinkers in history. The simplest and earliest type of robot was a fixed sequence type. The development in robotics is towards adaptive robots having sensory abilities».

Let's form the logic-linguistic models for each of the sentences of preset text (formal representation and model with substitution specific words) [8].

$$P_1 \,\&\, P_2\left(x_1, c_1, x_2[x_3[x_4\{c_{41}\}]]\right), \qquad (1)$$

where P_1&P_2 – predicate; x_1 – predicate variable subject; c_1 – predicate constant; x_2, x_3, x_4 – predicate variable arguments; c_{41} – predicate constant, that indicates the characteristic of argument x_4.

$$Can \,\&\, tell\begin{pmatrix} robots,\ also,\ difference \\ [between\ [temperatures\{two\}]] \end{pmatrix}, \qquad (1')$$

$$P_1' \,\&\, P_2'\{c_{21}'\}\begin{pmatrix} x_1', x_2'\{c_{21}'\}, \\ x_3'\{c_{31}' \,\&\, c_{32}'\}[x_4'[x_5']], x_6' \end{pmatrix}, \qquad (2)$$

$$Is \,\&\, state\{sovereign\}$$

$$\begin{pmatrix} Ukraine, territory\{its\}, bodies\{high \,\&\, local\} \\ [state[power]], government \end{pmatrix}, \qquad (2')$$

$$P''\left(x_1'', x_2''[x_3''\{c_{31}''\}[x_4'']] \,\&\, x_2''[x_5''\{c_{51}''\}[x_4'']]\right), \quad (3)$$

$$Was\begin{pmatrix} he,\ one[scientists\{greatest\} \\ [history]]\,\& \\ one[thin\ \ker s\{greatest\}[history]] \end{pmatrix}, \quad (3')$$

$$P'''\left(x_{11}''' \,\&\, x_{12}'''\{c_{11}''' \,\&\, c_{12}'''\}, x_2'''[x_3'''[x_4''']]\right), \quad (4)$$

$$Was\begin{pmatrix} type \,\&\, robot\,\{simplest \,\&\, earliest\}, \\ fixed[sequence[type]] \end{pmatrix}. \quad (4')$$

The obtained models allow you to compare predicates and subjects of sentences without searching keywords. For example, model (3) and (4) have the same predicate, but the subjects have completely different semantic value, and therefore can't be the same in content. The subject of model (2) is word «Ukraine», predicate variables (arguments) of this sentence do not overlap in meaning with the subjects and objects of logic-linguistic model (1), (3), (4). This can be checked by using electronic semantic dictionary. Comparison of the same subjects of model (1) and (4) reveals that the words «robot» and «robots» - are nouns used in the singular and plural respectively. Predicate constant «also» informs that in text should be mentioned about the subject of logic-linguistic model earlier in this text. Thus, a sentence that is described by logic-linguistic model (4) should precede sentence (1). The remaining sentences are unrelated by content. Such conclusions were made by comparison the main components of the elementary logic-linguistic models of text information. If we complicate the comparison criteria and selection algorithm we can improve the comparative analysis of logic-linguistic patterns of natural language sentences.

Thus, we can make a permutation of sentences in preset text: «The simplest and earliest type of robot was a fixed sequence type. Robots can also tell the difference between two temperatures. Ukraine is a sovereign state with its territory, high and local bodies of state power, government. He was one of the greatest scientists and thinkers in history».

Conclusion

Semantic structuring of natural language sentences in text is not possible without the implementation of semantic analysis. The results of this analysis can be present as a semantic graph, FTN and in the form of logic-linguistic models. Researches demonstrate that the FTN can visualize relations between words in natural language sentences, but it is not easy in use. This is due to a variety of sentence structures and the number of used words. Speaking about relations between sentences in the text, the transition networks are not designed to handle large amounts of information.

The logic-linguistic models are able to display semantic-syntactic relations in natural language sentences. A detailed study of its components (predicates, subjects and objects), comparison and also application of synonymic dictionary allow to determinate common content components. Due to logic-linguistic models of text information it is possible to trace semantic relations between sentences and structure them in a document.

References

[1] V. Sh. Rubashkin, Predstavlenie i analis smusla v intellektualnukh informatsionnux sistemax [Presentation and analysis of the meaning in intelligent information systems], M., Science, 1989, 192 p.

[2] V. A. Zvegintsev, Novoe v zarubezhnoy lingvistike. Lingvistichaskaya semantika [New in foreign linguistics. Linguistic semantics], Vol. 10, M., Progress, 1981, 568 p.

[3] M. A. Krongauz, Semantika [Semantic], M., Publishing Center "Academy", 2005, 352 p.

[4] E. B. Hunt, Artificial Intelligence, Academic Press, Inc., 1975, 560 p.

[5] N. Chomsky. Three models of language. Cyber collection, Vol. 2, 1961, pp. 81-92.

[6] R. G. Buharaev, D. S. Sulaymanov, Semanticheskiy analis v voprosno-otvetnukh sistemakh [Semantic analysis in question-answer systems, Kazan, Publishing of university of Kazan, 1990, 124 p.

[7] Factor. South-Russian Research Center, content analysis. http://opros-center.info. [Online]. Available: http://opros-center.info/inform06.htm. [Accessed: Sep. 30, 2013]

[8] A. I. Vavilenkova. Logiko-lingvistuchna model yak zasib vidobrazhennya suntaksuchnukh osobluvostey tekstovoi informatsii [Logic-linguistic model as a way to express the syntax of text information], Mathematical machines and systems, Institute of Mathematical Machines and System Problems of the National Academy of Sciences of Ukraine, 2010, pp. 134-137.