Science and Technology March 2013, Special Issue: 41-47
DOI: 10.5923/s.scit.201301.07

The Mechanism Analysis of Natural Language Texts in Order to Construct A Model of the Full-text Document

A.S. Lebedev

Computer Engineering Dept., Zaporozhye Institute of Economics and Information Technologies, Zaporozhye, 69041, Ukraine

Abstract The article describes the main methods of analyzing natural language. Isolated linguistic text analysis method and the basic steps of text analysis. Detailed itemized mechanisms of morphological and syntactic analysis. Considered the stages of the formation of a semantic network. Created a software system which produces a semantic network of the source text in natural language.

Keywords InformatiCS, Knowledge Base, Antiplagiat System

1. Introduction

In modern society the important role played by information technology. Over time, their importance is continuously increasing. But the development of information technology is very uneven: if the current level of computing and communication is amazing, in the field of semantic information processing is much more modest success. These successes are dependent primarily on advances in the study of the processes of human thought processes of verbal communication between people and the ability to model these processes on a computer.

When it comes to creating advanced information technology, the problem of automatic processing of textual information presented in natural languages, come to the fore. It is determined that the person's thinking is closely linked with his language. Furthermore, natural language is a tool of thought. He is also a universal means of communication between people - a means of perception, storing, processing and transmitting information.

To address the objectives necessary to move to a new level of processing and presenting information. Specifically, the processing of natural language is necessary to build a structure that will, to some extent formalized document in natural language, to build its model. Using these approaches, receiving document model, we can achieve universal search queries or to create a universal and unique knowledge base.

Note, however, that neither the one nor the other is not an end in itself, it is more of the components of more complex problems or support mechanisms for software systems with a very different purpose.

* Corresponding author:
bshdrago@yandex.ru (A.S. Lebedev)
Published online at http://journal.sapub.org/scit

For example, when using the considered approach, you can create a search engine of new generation, which will search for information is not a keyword, and within the meaning of the required information.

You can design a system for automatic text generation, which will generate a sample of data from a vast array of knowledge in a short text, and it is possible to set not only the subject and scope, and complexity of the method and level of generated text.

Actual development of the interpretation of texts. If a model is developed showing the meaning of the text, that meaning can be projected to any natural language. That is, the system is able to produce a translation of the text on the same level at which it would have made a professional translator with the subject matter and level of difficulty of the text.

Relevant for today and the so-called antiplagiat system. These are programs that allow you to compare some text to a pre-knowledge base, and determine the percentage of plagiarism in the text. Moreover, such systems need to compare the basic semantic relations, and not perform a comparison on mutual entry into documents and phrases.

To address the problem of processing and presentation of the text has been developed further described system. In order to more applied work, then, in this article, the system development will be considered in the spectrum antiplagiat systems. In addition, given that the system is designed for the CIS, the examples discussed in the article are, in Russian. However, the principles of this system are applicable to any language

2. Development Tools

Given the wide range of applications of the idea, the fundamental difference in the development tools were not. However, given that the spectrum atiplagiat systems more

appropriate site is to write the whole system was set up as an independent Internet page.

As the existence of the original base was chosen package Denver

The contents of the basic package of Denver:

Apache - is HTTP-server. That is the basis of the Apache package.

- SSL - is a secure protocol that allows you to communicate important information in a safe manner. These protocols are often common, when used on the page is very important operations (for example, WebMoney).
- SSI - a language that lets you create and work with pages shtml.
- PHP5 c different modules (mod_rewrite, mod_php).
- MySQL - a relational DBMS. All new sites have long been all of the content stored in the database.
- PhpMyAdmin - this is a common web-based application that is responsible for the administration of MySQL.
- SendMail - emulator SMTP-server, it is possible to write a feedback form and test the results by sending an email.

Pros and Cons:

- [+] The small size and portability.
- [+ / -] The absence of unnecessary modules.
- [+] Autonomy. That is the registry, system folders intact. You can run from a flash drive. Not needed deistallyator.
- [-] No external access, that is, your friends can not come and see your work.
- [-] Security

Denver - it's a simple local server main tasks - checking of scripts, work with MySQL, work with the emulator SMTP.

As the development environment was used ZendStudio 7.2.0. Zend - professional environment for developing and debugging web projects. A distinctive feature of the program is the ability to do remote debugging and profiling. Note that for remote debugging is required to install ZendStudioServer, which is a server module. The composition ZendStudio as an option package included PHP with an extensive list of pre-compiled extensions. If you already have pre-configured version of PHP, ZendStudio product easy to integrate with it. Also in the package environment is ZendOptimizer, a server module to run encoded using ZendEncoder ZendSafeGuardSuite and scripts as well as some of their speeding. The seventh version ZendStudio adds support PHP 5.3, integration with ZendFramework and ZendServer, improved editing of source code and various performance improvements program.

The main features of the development environment is:

- Integratsiy as ZendiZendServerFramework
- Analysis of the code and a quick fix
- Quickly create a new file
- Support for PHP 4.x and 5.x
- SyntaxColoring
- Using code templates (PHP, PHPDoc, New File)
- Detection of errors in real time
- Using Book marks
- Internal Browser
- Commenting on the PHP code
- Search for text and elements of PHP code
- Search and replace text in files
- Integrated TODO mechanism
- Support for HTML and CSS
- Debugging PHP code
- Using toolbars in IE and Firefox

3. The structure of the Developed System

Schematic diagram of the system is simple at first glance, but in fact close to the structure of programming languages, compilers, thus increasing the accuracy and quality of the analysis.

The algorithm of the method can be divided into four main sections:

- Thedivision of the text
- Analysis of the words
- Review of the proposal
- Merge to form a semantic network.

At the stage of the construction of a semantic network was offered a special mechanism to handle synonyms and homonyms, the replacement of set phrases and phraseology. Because the system considers the entire text as a collection of some of the identifier, such as synonyms, even in the dictionary have been the same identifier. In the case of the result of a disconnected graph suggested checking for associative relationships. It is possible to create connection "it is" between words that have no direct correspondence. Proceed to a detailed description of each block (Fig. 1).

The division of the text:

Since any natural language in terms of a computer is a consistent set of characters, for text analysis, you must first share the suggestions and make words in sentences.

To perform this task, the simplest, and so, and fast, the method is the use of regular expressions. Clearly, any proposal ends with a special character from which there can be either the following sentence, or the end of the text. A prior proposal or beginning of the text, or a special symbol of the end of previous proposals. Consequently, we can say that all is between the beginning of the text or special characters and other special characters have a suggestion.

In the same way the word should be considered all that lies between the beginning of the text, or space, and special characters or spaces. However, it should immediately distinguish different word sentences, that did not happen in the future confusion.[1]

Analysis of the words:

For this process, the key factors are speed and accuracy, however, given the rising trend of hardware capacity, as the key factor was chosen precisely characterize speech. Therefore, as the primary method for determining the chosen vocabulary methods.

At the same time, given the number of words in the Russian language and the amount of space on the media that these words would have taken, had to abandon the store

words in the whole form. In the sense that the words with all possible endings, suffixes, prefixes, and so on can produce in large numbers, and to find in a dictionary search word is difficult and long. Instead, a mechanism that allows you to store media on only parts of speech, with fully determine all the necessary characteristics of speech.[2]

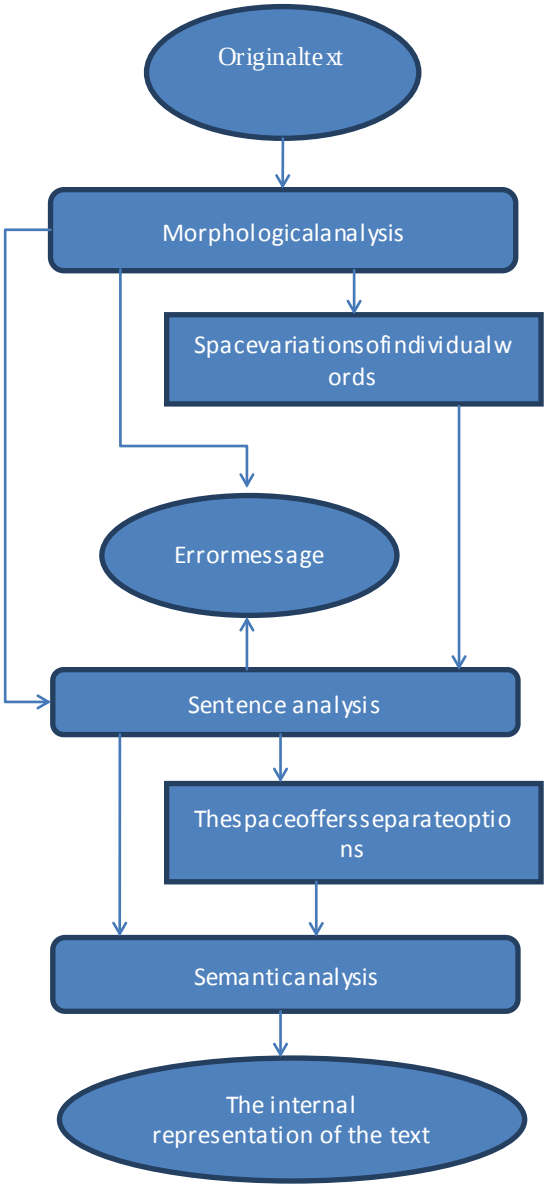

Figure 1. Outline of a draft

This storage option requires special approaches to the search and selection of new words.

Sequence analysis:

- Check for the presence of the word in its base of the main structures in the form of general (often it is the roots of words, but may make some additional values). Simple words, consisting only of the root can not be divided, therefore, they are stored in the database as a whole, together with their characteristics.
- If the word is not found, make check for the presence of the word endings. The number of words containing in the structure only the root structure and the end is small, but the definition of the presence of suffixes without preliminary endings possible.
- After that searches for key words in the database structures with pre-separation ends.
- If the word is found, it is necessary to determine the characteristics of the newly formed and return result. The result is a division of word variants and all variants of the characteristics of speech. Characterization is necessary to consider the impact of the end of the word, namely its case.
- If the word is not found necessary to check for the presence of the word suffixes. A check for the presence of suffixes, by comparing the suffixes from the database to the latest word characters, the root of a pre-separated.
- search for words in the base of the main structures, the suffix previously separated.
- If the word is found, the consistent definition of its characteristics. Suffix imposes a special mark on the characteristics of the word, because it can change all data at once. For example, the suffix "л" forms of verbal constructions neuter nouns, and adjectives. That is, forms the ambiguity. The point in this issue puts the ending, which has different adjectives and nouns.
- If the word is not found, you need to check the presence of a nested suffix. It should, in fact, cut the suffix and make checking for new suffixes. So, will check for the presence of all possible suffix, regardless of their number.
- When checking for suffix did not give the desired result, that is, the word was not defined in any of the options, you need to inspect for the presence of the prefix. To facilitate the work of the mechanism and its implementation prefixes and affixes taken as equal. Due to the fact that prefixing method of forming words is the smallest number of words on it was combined with the suffix-prefixed method, given that the suffix can be null.
- Search for words in the base of the main structures, the prior removal of all pre-defined structures.
- If the word is defined, it is necessary to identify the characteristics (prefix imposes only a shade of meaning to the word but does not affect its performance) and outstanding performance.
- If the word is not defined, you need to inspect for the presence of suffixes, after removing all the installed construction.
- Search for words in the base of the main structures, pre-separating all predefined structure.
- If the word is defined, to determine its characteristics with the prefixes, suffixes and endings and produce the output.
- If the word is not defined, to inspect for the presence of the words of another prefix, by repeating the process, before separating already defined prefix.

The fact that the word was not defined in any of the above cases, means that at least one of his designs is not in the knowledge base. Therefore, you must use a mechanism to at least approximate the characteristics of the word - naive method definition. Work of the method is based on the analysis of the identified structures, which have their own

characteristics.

Sequence analysis:

• Perform a review of all possible endings. Since the word is not defined, then we can certainly say that the word belongs to the Russian language, and therefore it is logical to assume that, in accordance with the termination can be defined and characteristics of speech.

• Identify the characteristics of endings, take them as possible alternatives to the characteristics of speech.

• The same actions apply to the suffix and prefix.

• Outstanding results.

Both of these methods are needed and are designed to operate in parallel, in order to complement each other in case of need.

The result of this block is an array, each element of which contains information about each word:

• The word

• An array containing all the options cut the text, namely:

o All to kens for the cutting

o All specifications for the cutting

Such a structure can provide ambiguous results. Also check all the options parsing words.

Analysis of the proposal:

The input of this block is fed a set of tuples of words with cutting characteristics, based on these characteristics to make connections between words in sentences. However, given that the analysis produced ambiguous words, there may be errors in the analysis of the level of analysis of the proposal. To do this, once it is necessary to establish the sequence analysis and selection of the correct result.

Sequence analysis:

• Select a sample of one of the sequences, that is, of all the options to choose the parsing of all the words one by one.

• Perform analysis of the selected sequence by the rules of sentence structure.

• If the analysis is not all words are involved in a bunch of proposals, such variant sequences are considered incorrect and the results of discard. Then return to the beginning.

• If the sequence is first obtained, then remember it, or compare the links and to establish the most stable.

• Based on the analysis and obtained stable ties with the rules, to determine the most probable parse the words, and thus the proposal.

The most important point for the block is a set of rules and consistency of their application. Rules themselves are selected in such a way as to fully and accurately establish links in a sentence without using any additional information.

At the output of the tree we offer. Obtained at the output of this stage trees are inputs of the next block.

Merging to form a semantic network:

At this stage, taken as a basis of the fact that in the first part of the text, the so-called intonation, formed the main concepts and objects of the text. It follows that, consistently integrating all the proposals into a single structure from beginning to end, you can get the full meaning of the text.

As the merge points to take the basic objects and actions, as they carry the main display of the text, and the rest, for the most part, reflect shades.

However, there are several problems associated with the features of the language, namely, the presence of synonyms, homonyms, simple pronouns and most unpredictable - associative relationships. To solve these problems, you should create multiple thesauri.

The procedure for the formation of a semantic network:

• Take one sentence

• Elimination of insignificant words with maintaining the integrity constraints that eliminate prepositions, conjunctions, and so on.

• To remedy the pronouns at the sentence level, a word which is replaced by a pronoun in the sentence has already occurred, it is used pronouns in the same number in the same way.

• Replace structures with the same meaning.. In a specially constructed thesaurus store design, have the same meanings as well as sets of relationships in which they can enter special codes and replacement. Thus, we can eliminate not only synonyms, homonyms, and set phrases and expressions. Such substitution will significantly increase the percentage of compatible merge proposals, and to avoid the appearance of false merger. Because some phrases may contain smaller structures to be replaced and part of a larger, necessary to test for the presence and replacement cycles, until all possibilities are exhausted.

• Replacing pronouns at the level of the text. If a replacement is at the sentence level is made, then one of the previous proposals should be a noun, which was replaced. It stands in the same number and gender as the pronoun. Consistently passing the word back, replaces the nearest found the word which corresponds to the parameters.

• Merge all the semantic relations on objects and actions. If the proposal first, then merge with nothing, so it is easy to make no change.

• If the merging is not possible, the text is stored as a set of blocks for as long as all possible points of the merger will not be installed.

• Once it is established that the text can not be connected using standard methods, it is necessary to analyze the presence of associative relationships. That is, the relationships between words that are not synonymous, and often outside the text does not have any relationship with each other or relationships, but in the text they can form a set of constraints (eg, "it is", "part-whole"). Establish these relationships allows special thesaurus containing possible links.

Because the text is connected, then the establishment of a coherent sense of the text display, if all the above conditions are necessary.

4. The System

The simplest version of a semantic network - Display meaning of one sentence. In the simplest case consists of subject and predicate, they are usually applied at least one slave word. For example:

«Веникстоитвуглу.»

In this case, the semantic network contains subordinate word "углу" associated with the preposition "в". In the semantic network prepositions and conjunctions are removed as insignificant.

For an easier perception, this article IDs words are replaced by their text counterparts. A connection of any type are displayed as arrows.

Consequently, the semantic web will be as follows:

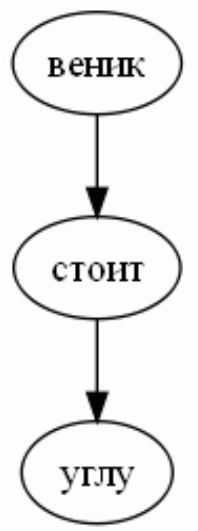

Figure 2. example of a simple semantic network with one word slave

Slightly more complex example illustrates not intelligibility of the system to analyze more complex examples

«Он подобрал выброшенную бумажку и положил ее за пазуху.»

This semantic network is as follows:

Figure 3. An example of a semantic network composite offers

The actual result of the system is as follows

Table 1. The result of the system

	0	**1**	**2**	**3**	**4**	**5**	**6**
0	0	1	0	0	1	0	0
1	0	0	0	1	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	1	1
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0

where:

0 - он

1 - подобрал

2 - выброшенную

3 - бумажку

4 - положил

5 - ее

6 - пазуху

And according to the intersection is 0 - no connection, 1 - the presence and direction of communication.

Two of the links (presence and absence) have been taken to simplify the illustration. In reality, communication and their variants are established based on the rules for constructing sentences

5. Discussion of Results

In the course of the work was the comparative analysis of modern approaches to building semantic models of test documents. The analysis phase identified the strengths and weaknesses approaches examined.

Considered such systems as:

- System SnePS

The basis for this method was the idea in the text representation of objects as instances of the class. All knowledge about individual objects, classes of objects, their properties, abstraction, action, propositions, rules and meta-rules is represented as a network of nodes and directed named relations.

Node in the network belongs to one of four types: basic, alternating, and a molecular model - and a unique identifier. The base unit for the element is conceptually different from any other node. Variables nodes correspond to arbitrary elements of the network and can be identified with some of the basic unit. Of base and variable components of the arc can not go, their properties are defined by arcs belonging to them.

Molecular nodes correspond to propositions, models - functional term with free variables and are used to search the network.[3] The system is widely used as a means of developing pilot applications using natural language, as it includes the basic mechanism for representation and automatic construction of semantic networks with minimal structure and inference on these networks.

The close connection of this system to the tasks of natural

language processing due to the fact that its architecture is focused not on manually creating a semantic network, and in its construction as a result of extraction of knowledge from various sources, mostly natural language text.

• System SNOOP

SNOOP system uses a mechanism that integrates network representation with object-oriented programming and production rules.

Network nodes and the relationship between them belong to the class described by a separate network of inheritance, in which a class is determined by a set of possible fields and properties associated with the object corresponding to the class.

Classes describe the properties of objects by means of groups of production rules, each of which consists of a pattern, you can navigate through the network, the terms of the nodes found pattern, and action to change the internal structure of the nodes and network changes.[3]

There are two kinds of relationships inherit a more general class of communication actantial Communications, and pronoun class inherits from the noun. Rectangles and directed arrows indicate nodes and relationships in the network.

Another specific feature of this formalism is the mechanism of the modularization of production rules, separated on the properties that are associated with the objects of the corresponding classes and consists of a set of production rules. This formalism was used for analysis of texts in a narrow domain.

• Linguistic Processor LP Krisin for complex information systems

Со стороны своей внутренней структуры данная система представляет собой многоуровневый преобразователь. В нем различаются три уровня разного представления текста:

- Morphological
- Syntax
- Semantic.

Each of the sub-levels served by one of the components of the model, by arrays of rules, a dictionary or dictionaries. At each level is formed formally called morphological, syntactic and semantic structure, respectively. Thus, we can talk about the rigid structure of technology.

Morphological structure is a set of words, their parts of speech and basic characteristics.

The syntax is, in fact, a tree proposal, which reflects the structure of hard sentences.

The semantic structure of a graph that reflects the meaning of the text, its structure is combined with the proposals.

This mechanism reflects the structure of the text in the form of a set of objects. These objects, in general, are presented as a set of attributes describing it - a statement about the object and the identifier of the object may be a serial number or a combination of statements, the so-called "core" characteristics that distinguish it from many of the other objects.

Serial number of statements represents a concatenation of all features of the object, and it can be seen as the links between them. All these objects are stored in a table. The table rows - the links between features, columns - name tag, and their intersection - the value attribute. Then each row corresponds to a statement about a single object.

It should be noted that, in this view the structure, the number of fields in the records may be constant or vary from record to record, which causes some difficulties in the further analysis. Namely, in the records of the functional role of a permanent format of concepts expressed by positional means (through the establishment of a specific point in each field), and in the records of variable format - using special code combinations (keywords, indexes role, etc.).

It should also be noted that this mechanism involves the use of the analogy. This method, in turn, does not guarantee a correct analysis, but can be used if other methods of analysis for success.[4]

According to the results of the analysis as a basis for further research was chosen approach in the language processor LP Krysin.

A characteristic feature of the method proposed by the author is the mechanism analysis of the word.

Suggest my own dictionary structure and methods of work with him.

Identified specific mechanisms of construction of a semantic network, avoiding ambiguity and solve possible disconnected graph semantic network.

For comparison site antiplagiat.ru were loaded two texts:

• AdobeFlash (MacromediaFlash), или просто Flash— мультимедийная платформа компании Adobeиспользуется для создания веб-приложений или мультимедийных презентаций. Широко используется для создания рекламных баннеров, анимации, игр, а также воспроизведения на веб-страницах видео- и аудиозаписей. (19% of original)

• AdobeFlash (MacromediaFlash), или просто Flash— мультимедийная основа компании Adobeиспользуется для разработки веб-приложений или мультимедийных показов. Широко используется для разработки рекламных баннеров, анимации, игр, а также воспроизведения на веб-страницах видео- и аудиозаписей. (100% оригинальности)

The reason for this variation is that the site does not take into account the presence of synonyms.

As mentioned previously developed system stores the words synonymous with the same codes. As a consequence, both the text will be identified as identical in meaning. A kind blagodaryaraznogo rules for proposals permutation will not help hide the authorship of the text

6. Conclusions

The system which is described has a complex architecture. However, this system has several advantages for the analysis of natural language texts:

• Dedicated stages of text analysis are clear limits of that

allow you to distinguish the stages of the system and reduce errors

- Morphological analysis is a series of tests, each of which is based on the previous one, which ensures the absence of ring bonds
- Parsing involves not just the establishment of the fact that what we offer is a particular structure, but allows us to analyze all possible options for the analysis of words and sentences
- Semantic analysis has a number of tools able to deal with the problem of substitutes words, phraseology, and other things.

REFERENCES

[1] D. Koterov, A. Kostarev . PHP 5. - BHV-Petersburg, 2006. - 1120p.

[2] Online Available: N. S. Valgina. Modern Russian language. - Hi-edu.ru.

[3] Online Available: As a means of ontology mapping knowledge. - Recordmetri.ru.

[4] L. P. Krysin. Language processor for complex information systems. - Spb.: Moscow, 1992. - 256s.

[5] Online Available: I. P. Kuznetsov. Linguistic and algorithmic aspects of selecting objects and relationships from domain-specific texts. - Dialog.ru.

[6] Online Available: A. Ogarov. Reporting mechanisms meaning of the text. - webinfo.ru

[7] Online Available: V. I. Litvinin. Modeling knowledge of the world based on rhematic graphs. - aienkod.ru

[8] M. Shevchenko. Technology of multiversion object-oriented structure of the text: Dis. Candidate. Physics and Mathematics. Sciences: 05.13.11 MOSCOW STATE UNIVERSITY. - K., 2005. - 212 p.