Journal of Theoretical and Applied Information Technology
10th April 2016. Vol.86. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ESTIMATION OF THE DEGREE OF SIMILARITY OF SENTENCES IN A NATURAL LANGUAGE BASED ON USING THE LINK GRAMMAR PARSER PROGRAM SYSTEM

YERIMBETOVA A.S.[1], MURZIN F.A.[2], BATURA T.V.[2], SAGNAYEVA S.K.[1], SEMICH D.F.[2], BAKIYEVA A.M.[3]

[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
[2]A.P. Ershov Institute of Informatics Systems, RAS, Novosibirsk, Russian Federation
[3]Novosibirsk State University, Novosibirsk, Russian Federation

E-mails: aigerian@mail.ru, murzin@iis.nsk.su, tatiana.v.batura@gmail.com, sagnaeva_tar@mail.ru, deiman32@ngs.ru, m_aigerim0707@mail.ru

ABSTRACT

Our main goal is to construct the algorithms that can estimate the document relevance on the basis of the text structure analysis. It is important that this estimate should be based on the context of the search query and not limited only by keywords, their similarity or frequency. The semantic-syntactical relations between words built by the program system Link Grammar Parser (developed in Carnegie Mellon University, USA) can be used to solve these problems. They allow us to develop the methods of comparison of the sentences in a natural language and introduce certain measures of the closeness (similarity) between the sentences. These measures take into account both lexical and syntactic relations between words. Experiments with different types of sentences and links took almost one year. It was observed that there is no need to use too many links. First, the use of some links leads us to the analysis of diagrams which correspond badly to intuition and principles of classical linguistics, and it is not clear what we can do with them further. Second, there is also a complexity aspect. If there are fewer links, the algorithm works faster. Therefore, a compromise is necessary. The minimum variant giving good results is when only eight connectors (links) are used. One of the problems we solve in the current time is the development of a parser like Link Grammar Parser for Turkic languages most frequent in the Internet, such as Kazakh, Uzbek (Cyrillic and Roman alphabets), and Turkish. The results of our research are planned to be used in different information retrieval systems.

Keywords: *Natural Language Processing, Syntactic Analysis, Link Grammar Parser, Relevance, Turkic Languages.*

1. INTRODUCTION

Under the conditions of rapid growth of the volume of information resources, it is required to improve the quality of information retrieval. Many researchers [1, 2] consider deep semantic analysis of texts necessary for making the semantic images of texts which can be the basis of fine ranking of documents. This approach, undoubtedly, is the most reasonable; however, it requires a careful and long-term work on the creation of suitable tools for automatic text processing [3]. In particular, the detailed description of various fields of knowledge is required. Therefore, a search for partial solutions, one of which is presented in this paper, is also useful.

Our main goal is to construct the algorithms that can estimate the document relevance on the basis of the text structure analysis. It is important that this estimate should be based on the context of the search query and not limited only by keywords, their similarity or frequency.

The semantic-syntactical relations between words built by Link Grammar Parser can be used to solve these problems [4, 5]. The basic algorithm for calculating the degree of correspondence between link diagrams and natural language constructions is described in [6–8]. The studies were completely focused on the English language sources. Based on the above mentioned ideas, the "iNetSerch" information retrieval system was implemented. Testing has shown that the proposed

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

algorithm efficiently solve the problems of information retrieval. The methods which generalize the approach used in the basic algorithm and take paraphrases into account are presented below.

But from considered experiments, it is possible to make a conclusion that further development of this method will not lead to substantial improvement of the obtained results. One of the reasons is that the possibilities of Link Grammar Parser at the current stage of work are almost completely exhausted. And, in spite of the fact that Link Grammar Parser possesses a number of advantages (high speed, partial coverage of semantics, many examples of its successful application in the systems of Internet texts filtration), it makes us to stay at the level of syntax with partial semantics coverage. Therefore, if we want to have essential advancement, it is necessary to move to a higher level of knowledge engineering.

2. THE LINK GRAMMAR PARSER SOFTWARE SYSTEM

Link Grammar Parser is a syntactic analyzer of the English language developed in 1990th at the Carnegie Mellon University, USA. Note that, in general, the underlying theory differs from the classical theory of syntax. Having received a sentence, the system attributes it with a syntactic structure which consists of a set of marked links connecting the pairs of words. The detailed description of the system can be found in [4, 5].

Link Grammar Parser includes approximately 60000 dictionary forms. It allows us to analyze a huge part of syntactic constructions, including numerous rare expressions and idioms. The parser work is stable; it can skip a part of a sentence it cannot understand and define some structure for the rest of the sentence. It is capable of processing an unknown lexicon and doing reasonable assumptions about the syntactic category of unknown words based on the context and writing. The parser contains data about various names, numerical expressions, and punctuation marks.

The rules of words connection are described in the set of dictionaries. For each word in a dictionary it is determined by means of what connectors it can be connected with other words of a sentence. A connector has a name with which the considered unit (word) can enter a sentence. For example, the mark S corresponds to communication between a subject and a predicate, O is a connector between an object and a predicate. There are more than one hundred most important basic connectors. To denote the direction of a connector, the sign "+"is used to indicate the right connector and the sign "–" to indicate the left connector. Left-directed and right-directed connectors of the same type (see Figure 1) make up a connection (link).

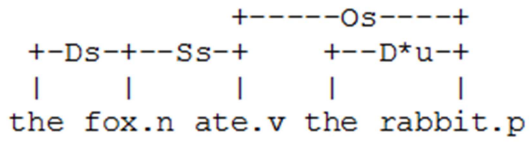

Figure 1. An Example of Syntactic Analysis of a Sentence

The obtained diagrams, as a matter of fact, are analogues to the so-called trees of submission of sentences. In the trees of submission, it is possible to raise a question from the main word in the sentence to the minor one. Thus, words are built in a treelike structure. The syntactic analyzer can give out two or more diagrams of analysis of the same sentence. This phenomenon is called syntactic synonymy.

The main reason why the analyzer is called a semantic system is the unique set of connectors (about 100 basic ones, and some of them have three or four variants). In some cases, the authors of the system pass on to almost semantic classifications constructed exclusively on syntactic principles.

For example, the following classes of English adverbs are allocated in the system: situational adverbs concerning the whole sentence (clausal adverbs); time adverbs; introductory adverbs which stand in the beginning of the sentence and are separated by comma (openers); the adverbs modifying adjectives, etc. As for the advantages of the system, it is necessary to notice that the procedure of finding the variants of the syntactic representation is organized very effectively. The process of construction is not top-down or bottom-up, but all the hypotheses about the relations are considered simultaneously: at first all possible connections are constructed by dictionary formulas, and then the possible subsets of these communications are allocated.

Of course, it leads to some algorithmic opacity of the system, because it is very difficult to track

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

all relations at once. Secondly, it leads not to a linear dependence of the speed of the algorithm on the number of words, but to exponential one, because the set of all variants of syntactic structures of the sentence containing N words in the worst case is equipotent to the set of all spanning trees of the full graph with N nodes.

The last feature of the algorithm forces the developers to use a timer to stop the procedure which works too long. However, all these lacks are compensated by linguistic transparency of the system in which rather simple valences of words may be registered, and the order of gathering the valences in the algorithm is not strictly fixed, i.e. the connections are constructed simultaneously, which completely corresponds to our linguistic intuition.

Let's consider also the negative moments.

1. The practical testing of the system shows that during the analysis of complicated sentences of length more than 25–30 words, a combinatory explosion is possible, and in this case the result of the analyzer work is the "panic" graph, which as a rule has several variants of syntactic structures, which is inadequate from the linguistic point of view.

2. The application of the ideas described above is complicated for inflective languages, such as the Russian language, in view of the considerably increasing volume of dictionaries because of the morphological complexity of the inflective languages. Each morphological form should be described by a separate formula, where the bottom index of a connector name should provide a coordination procedure. This leads to an increasing number of connectors. For agglutinative languages (for example, Turkic), the system becomes even more complicated.

3. THE BASIC ALGORITHM OF THE COMPARISON OF SENTENCES

Below we assume that two sentences $\bar{x} =< x_1,\ldots,x_n >, \bar{y} =< y_1,\ldots,y_m >$ are given, i.e. the sentences are considered as vectors with words as their components. We suppose that their analysis is made by means of the Link Grammar Parser system. Let's consider the set of all pairs $< i_1, i_2 >, < j_1, j_2 >$ such that the words x_{i_1}, x_{i_2} and y_{j_1}, y_{j_2} are connected by links of the same type. Thereby the words x_{i_1}, y_{j_1} and x_{i_2}, y_{j_2} are close according to some criterion, for example, their normalized forms are identical, they are synonyms, words are similar by writing, etc. Some variability of the algorithm is possible here. Also, it is possible to ignore the auxiliary words: articles, unions, pretexts, interjections, etc. Let's assume now that I is a set of the pairs mentioned above and taken into consideration and its cardinality $|I| = n$.

Then, let n_1, n_2 be the numbers of the links obtained as the result of the analysis of the sentences $\bar{x}, \bar{y}$, respectively. As a measure of similarity of two sentences, it is possible to introduce $\mu_0(\bar{x},\bar{y}) = n/\max(n_1,n_2)$ or $\mu_1(\bar{x},\bar{y}) = 2n/(n_1+n_2)$. In the following section, the approach will be essentially generalized. It will be shown that the basic algorithm considers only the so-called invariant connectors, not taking into consideration more complicated logics.

Thus, the method described above allows us to introduce certain measures of the closeness (similarity) between sentences. These measures take into account both lexical and syntactic relations between words. The minimum variant giving good results is when only eight connectors (C, CC, S, SI, SF, SFI, SX, and SXI) are used.

Six links have been allocated that can dramatically aggravate the situation. Therefore it is useful to omit them. Approximately 45 connectors were analyzed.

Experiments with different types of sentences and links took almost one year. It was observed that there had been no need to use too many links. First, the use of some links leads us to the analysis of diagrams which correspond badly to intuition and principles of classical linguistics, and it is not clear what we can do with them further. Second, there is also a complexity aspect. If there are fewer links, the algorithm works faster. Therefore, a compromise is necessary.

Journal of Theoretical and Applied Information Technology
10th April 2016. Vol.86. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1 – The List of the Most Important Links of the Link Grammar Parser System

Link	Description
C	connects subordinating conjunctions, verbs or adjectives with the subjects of subordinated sentences
CC	is used to connect coordinating conjunctions
S	connects subject-nouns to verbs
SI	connects a subject to a verb in the sentences with the inversion of the principal parts of the sentence
SF	connects a subject expressed by "it" or "there" to a verb
SFI	connects a subject expressed by "it" or "there" to a verb in the interrogative sentence with the inversion of the principal parts of the sentence
SX	is used to connect the pronoun "I" to verbs "was" and "am"
SXI	is used to connect the pronoun "I" to verbs "was" and "am" in the cases of subject-verb permutation

4. LOGICAL METHODS OF EVALUATION OF THE SENTENCE SIMILARITY

As before, we suppose that L is a set of words in a natural language. For any word $x \in L$ we will denote its normalized form by $Norm(x)$. The formula $Syn(x, y)$ means that x, y are synonyms.

There are two forms of equivalence:

1) $x_1 \approx x_2 \leftrightarrow x_1 = x_2 \vee Syn(x_1, x_2)$

2) $x_1 \equiv x_2 \leftrightarrow Norm(x_1) = Norm(x_2)$

A sentence may be considered as a vector with words as its components, $\bar{x} =< x_1, \ldots, x_n >$. The function $Norm$ can be naturally extended onto sentences:
$Norm(\bar{x}) =< Norm(x_1), \ldots, Norm(x_n) >$.
The text $T =< \bar{x}_1, \ldots, \bar{x}_n >$ is a sequence of sentences.

Let the formula $\bar{x} \mid= P(x_i, x_j)$ mean that in the link diagram of the sentence $\bar{x} =< x_1, \ldots, x_n >$ obtained by Link Grammar Parser, there is a connector of the type P going from the word x_i to the word x_j. The sign $\mid=$ means that we consider a model. The basic set of the model is the set of pairs $\{< 1, x_1 >, \ldots, < n, x_n >\}$. As the same word can occur in the sentence two or more times, it is necessary to consider the pairs instead of separate words. This imply that $\bar{x} \mid= \varphi$, where φ is, for example, a formula of the first order logic, is a correct designation. Indeed, $\bar{x}$ is a designation for both a vector and a model at the same time.

Let's assume that two sentences are given: $\bar{x} =< x_1, \ldots, x_n >$ and $\bar{y} =< y_1, \ldots, y_m >$. It is interesting to consider the function f such that $dom(f) \subseteq \{1, \ldots, n\}$, $range(f) \subseteq \{1, \ldots, m\}$ with additional properties of the form: $f(i) = j \rightarrow x_i \approx y_j$ and $f(i) = j \rightarrow x_i \equiv y_j$, and others. When comparing two sentences or more, especially when performing the analysis of their similarity, verification of some logic properties is carried out. For example, let's consider $f(i_1) = j_1$, $f(i_2) = j_2$. The examples of such properties are given below.

1. The invariance of a connector
$\bar{x} \mid= P(x_{i_1}, x_{i_2}) \rightarrow \bar{y} \mid= P(y_{j_1}, y_{j_2})$

2. The replacement of a connector by a disjunction of others
$\bar{x} \mid= P(x_{i_1}, x_{i_2}) \rightarrow \bar{y} \mid= \bigvee_t Q_t(y_{j_1}, y_{j_2})$

3. The splitting a connector into two connectors
$\bar{x} \mid= P(x_{i_1}, x_{i_2}) \rightarrow \exists k\, (\bar{y} \mid= Q(y_{j_1}, y_k) \wedge R(y_k, y_{j_2}))$

4. The splitting a connector into two connectors by means of inversion
$\bar{x} \mid= P(x_{i_1}, x_{i_2}) \rightarrow \exists k\, (\bar{y} \mid= Q(y_{j_2}, y_k) \wedge R(y_k, y_{j_1}))$

Taking into consideration that $\bar{y}$ is a designation for a corresponding model, the third formula can be rewritten in the form
$\bar{x} \mid= P(x_{i_1}, x_{i_2}) \rightarrow \bar{y} \mid= \exists y Q(y_{j_1}, y) \wedge R(y, y_{j_2})$
. Analogously, the fourth formula can be written in a similar form.

The example of the analysis of two sentences, one of which is the paraphrased variant of another, is shown below (see Figure 2).

Journal of Theoretical and Applied Information Technology
10th April 2016. Vol.86. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

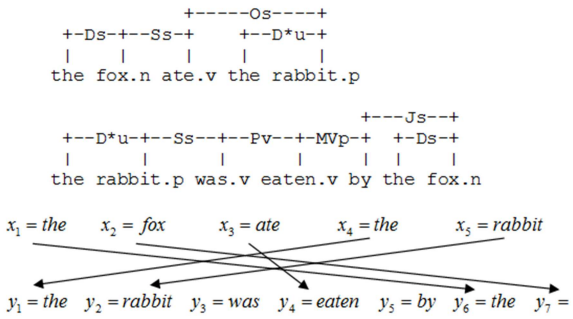

Figure 2. The Results Obtained by the Link Grammar Parser

Thus, we have

$$f(1)=6,\ f(2)=7,\ f(3)=4,\ f(4)=1,\ f(5)=2$$

As a result, we obtain:

1) $Norm(ate)=Norm(eaten)$ or $ate \equiv eaten$;

2) the connectors Ds and D*u remain, i.e. they are invariants;

3) $\bar{x}|=\mathrm{Ss}(fox,ate)\rightarrow \bar{y}|=\mathrm{MVp}(eaten,by)\wedge \mathrm{Js}(by,fox)$, i.e. there is splitting the connectors Ss with an inversion;

4) $\bar{x}|=\mathrm{Os}(ate,rabbit)\rightarrow \bar{y}|=\mathrm{Ss}(rabbit,was)\wedge \mathrm{Pv}(was,fox)$, i.e. there is splitting with inversion, but of another connector Os.

To summarize it is possible to say that there are rules of form $R_i:\bar{x}|=\varphi_i(x_1,x_2)\rightarrow \bar{y}|=\psi_i(y_1,y_2)$.

Further, function f is constructed and it is verified whether there are indexes $i_1, i_2,\ j_1=f(i_1),\ j_2=f(i_2)$ such that the rule R_i is satisfied on the concrete words from the sentences $\bar{x},\bar{y}$, i.e. $\bar{x}|=\varphi_i(x_{i_1},x_{i_2})\rightarrow \bar{y}|=\psi_i(y_{j_1},y_{j_2})$. For simplicity it is possible to say that the rule is satisfied by the pair $<i_1, i_2>$.

Let's consider a set of all such pairs $<i_1,i_2>$ by which one rule is satisfied. We denote this set by I , and its cardinality is $|I|=n$. Let's notice, that the Link Grammar Parser analyzer assumes the presence of only one connector between two words. Therefore, no more than one rule is satisfied. Let n_1, n_2 be the number of connectors obtained as a result of the analysis of the $\bar{x},\bar{y}$ sentences respectively. As a measure of similarity of two sentences it is possible to introduce $\mu_0(\bar{x},\bar{y})=n/\max(n_1,n_2)$ or $\mu_1(\bar{x},\bar{y})=2n/(n_1+n_2)$. This approach generalizes the approach used in the basic algorithm. More exactly, the basic algorithm takes into account only invariant connectors, not considering more complicated logics.

Let's consider an example of the similarity comparison of two sentences (see Figure 3).

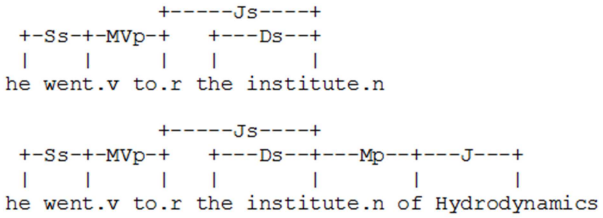

Figure 3. Comparison of Similarity of Two Sentences.

It is easy to see that $n_1=4,\ n_2=6$. Further we see that all the four connectors Ss, MVp, Ds, and Js from the first sentence remain (are invariant), therefore $n=4$. As a result we obtain $\mu_0(\bar{x},\bar{y})=4/\max(4,6)=4/6=2/3$ and $\mu_1(\bar{x},\bar{y})=2\cdot 4/(4+6)=8/10=4/5$. Thus we see that these measures of similarity are different.

For the English language we have 15 rules in addition to the three rules mentioned above. Thus, some of them allow three to five modifications. As a result approximately 30 rules may be used.

With respect to other languages, it is expedient to speak about the classes of languages. For example, the types of links and rules are practically identical for the Russian and Polish languages. In the Polish language in addition to six cases (the same as in Russian) there is an additional vocative case. Thus, it is possible to enter additional types of links. And for example, the types of links and rules are essentially different for the Russian and German languages. There are German constructions which are absent in Russian, but it is desirable to consider them.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

According to the morphological typology, there are analytic and synthetic languages. Synthetic languages are divided into agglutinative, fusional, and polysynthetic. Omitting the details, we say that the considered approach is most easily implemented for analytic languages, for example, English. The situation is more complicated for synthetic languages, in particular for flective (for example, Russian) and agglutinative (for example, Turkic) languages. There are two variants for these types of languages. The first variant is to use a small set of links that is enough for retrieval systems. The second, a more difficult variant is to use a large number of links. It is appropriate to use the second variant in translators. Polysynthetic languages include Paleo-Asiatic (for example, Chukchi and Eskimo) and some African languages. In this case, the situation is even more difficult, but the described approach is applicable.

Taking into consideration possible errors, it would be desirable to know how the algorithm itself will perform the analysis of similarity of such sentences: "the Fox eats rabbits" and "the Fox does not eat rabbits". Will the second sentence be considered equivalent to the first one?

It is a very interesting question how to differ automatically the positive and negative statements about the same thing. Omitting the details, we can say that the above two sentences will not be considered as similar. But if the sentences are long and the words in these sentences are the same except for their beginnings ("the Fox eats" and "the Fox does not eat"), then the proposed algorithm will identify these statements as equivalent. Certainly, it is possible to modify the formula for the evaluation of the similarity of sentences, for example, to assign a heavy weight to a link connecting a particle "not" with the verb entering the denominator of the formula. It is clear that further in-depth investigation is necessary.

To summarize, we have noticed that [9] and [10], where various measures of proximity between logical formulas are considered, have appreciably affected our research considered in this section.

5. SYNONYMY OF SYNTACTIC MODELS

The main problem of the automatic processing of texts in a natural language is that one word can have multiple meanings. This phenomenon is called polysemy. Conversely, one meaning may be expressed by means of various forms. In this case we speak about synonyms.

We use the following definition of syntactic synonyms. Syntactic synonyms are constructions which have identical or close semantic meaning, express similar syntactic relations and are able to replace each other in certain conditions [15]. The examples of syntactic synonyms are sentences from the previous section: *The fox ate the rabbit. — The rabbit was eaten by the fox.*

In other words, syntactic synonymy is realized in a transformation of syntactic units, for example, in extending a simple sentence. Extending a simple sentence can be described in terms of syntactic processes [16]: expansion, complication, deployment, combining, and joining.

Let's rewrite the syntactic processes mentioned above in a formal way using the notation introduced in the previous section.

1. Expansion is based on the fundamental property of grammar – the recursiveness. It consists in adding other elements of the same syntactic status and a common syntactic relation to a syntactic unit. For example: *I've known many ladies who were prettier than you. — I've known many ladies who were prettier than you are.*

Let's suppose that $\bar{x} =< x_1,\ldots,x_n >$ and $\bar{y} =< y_1,\ldots,y_m >$ are two sentences. At the same time, we consider them as models, and the corresponding predicates associated with the connectors of Link Grammar Parser are true in these models. Now the process of expansion can be written as follows:

$$\bar{x} \models P(x_{i_1},x_{i_2}) \rightarrow \exists k\left(\bar{y} \models Q_1(y_{j_1},y_k) \wedge Q_2(y_k,y_{j_2}) \wedge R(y_{j_1},y_{j_2})\right)$$

2. Complication can occur in a part of a predicate, i.e. in a verb phrase or an object.

2.1. For predicate, the complicating element will express the link with the subject. The second part of the predicate gets morphological structure of a non-predicative form. For example: John is expected to come to London. — John's coming to London is expected.

2.2. The complication of a direct object is possible after the verbs of specific semantics and it is reached by attaching an infinitive, a participle, an adjective, a predicative, or a prepositional phrase to a noun, pronoun, etc. as an object. For example: John saw his friend entering the hall. — John saw that his friend entered the hall.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The complication process can be written in the following form:

$$\bar{x} \models P(x_{i_1}, x_{i_2}) \rightarrow \exists k \left(\bar{y} \models Q_1(y_{j_2}, y_k) \wedge Q_2(y_k, y_{j_1})\right)$$

3. Deployment consists in a modification of an element of the sentence based on a link of syntactic dependency. As a result, a new syntactic construction appears, in which one component is syntactically dominating and others are syntactically dependent. For example, the sentence A boy put bottles can be extended to A nice little boy with rosy cheeks put three metal-topped bottles of milk quietly on my doorstep before seven o'clock as a result of application of the process of expansion to:

1) a boy (nice, little, with rosy cheeks);

2) put (quietly, on my doorstep, before seven o'clock);

3) bottles (three, metal-topped, of milk).

Then the deployment process can be written as follows:

$$\bar{x} \models P(x_{i_1}, x_{i_2}) \rightarrow \exists k, l \left(\bar{y} \models Q_1(y_{j_1}, y_k) \wedge Q_2(y_k, y_{k+1}) \wedge .. \right.$$
$$\left. (... \wedge Q_{l+1}(y_{k+l-1}, y_{k+l}) \wedge Q_{l+2}(y_{k+l}, y_{j_2}))\right)$$

It's worth noting, that expansion and complication (the processes earlier considered) are limited by the frameworks of a sentence part, i.e. they are similar to an internal transformation. In this aspect the deployment differs from aforementioned syntactic processes, because the results of the deployment are more complex syntactic units — word-combinations. Therefore, this process may be considered as external.

4. Combination of predicates in a sentence. As a result of superposition, a new part of a sentence emerges from two its parts. The predicates with meaningful verbs and their nominal part can be combined. For example: She looks out of the window and sees … — Looking out of the window, she sees …

The process of combining can be written as follows:

$$\exists k \left(\bar{x} \models P_1(x_{i_1}, x_k) \wedge P_1(x_k, x_{i_2})\right) \rightarrow \bar{y} \models Q(y_{j_1}, y_{j_2})$$

5. Attachment of a minor part of the sentence to a simple sentence by means of a coordinate conjunction, and thereby an attached part is not coordinated to any part of the basic sentence. For example: *Denis tried to escape, but in vain. — Denis tried to escape, but it was in vain.*

$$\bar{x} \models P(x_{i_1}, x_{i_2}) \rightarrow \exists k \left(\bar{y} \models Q_1(y_{j_1}, y_k) \wedge Q_2(y_k, y_{k+1})\right)$$
$$\left(\wedge Q_3(y_{k+1}, y_{k+2}) \wedge Q_4(y_{k+2}, y_{j_2})\right)$$

One can see that this formula is a special case of item 3 for $l = 2$.

As a result, we obtained **syntactic models** corresponding to the syntactic processes mentioned above.

Assume now that we have a text in the form of an ordered sequence of sentences $T = <\bar{x}_1, ..., \bar{x}_n>$. The **syntactic environment of the model** $\bar{x}_i$ is an ordered pair $<\bar{x}_{i-1}, \bar{x}_{i+1}>$, such that the sequence $\bar{x}_{i-1}\,\bar{x}_i\,\bar{x}_{i+1}$ is a part of the text T.

The most important and general criterion that allows us to establish the synonymy relation between syntactic models for many natural languages can be formulated as follows [17]: "syntactic models are synonymous, if they are interchangeable in an identical syntactic environment".

We obtain the following definition of the syntactic synonymy. Two syntactic models $\bar{y}_1, \bar{y}_2$ are synonymous, if $\forall\, \bar{x}_{i-1}, \bar{x}_{i+1} \left(\bar{x}_{i-1}\bar{y}_1\bar{x}_{i+1} \leftrightarrow \bar{x}_{i-1}\bar{y}_2\bar{x}_{i+1}\right)$.

6. LINK GRAMMAR PARSER FOR TURKIC LANGUAGES

One of the problems already solved is the development of a parser like Link Grammar Parser for Turkic languages most frequent in the Internet, such as Kazakh, Uzbek (Cyrillic and Roman alphabets), and Turkish. It should be noted that this kind of research was carried out by other authors [11–13].

The machine translation system from Kazakh into English and vice versa, using the link grammar and statistical approach, is considered in the paper by U.A. Tukeyev et al. [11]. Link Grammar plays an important role in the algorithm there proposed. The statistical approach is used for translation of polysemantic words. The developed models and algorithms have been implemented in the program of machine translation. According to the linguistic

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

classification, there are six different types of languages: SVO — Subject Verb Object; SOV — Subject Object Verb; VSO — Verb Subject Object, etc. These schemes reflect the typical structure of sentences. Turkic languages belong to the type SOV. A list of 13 links that naturally reflect the most important syntactic links between words in the sentences in the Kazakh language is described in [11]. It is important that the same links can be used in the development of parsers for other Turkic languages, due to the high degree of similarity not only of their syntax, but also the morphology and vocabulary.

In [12], the "statistical parser" of dependencies of the Turkish language is described, which is based on the statistical models of learning based on the sentences in the Turkish language from the so-called Turkish Dependency Treebank. As a result, the parser produces the dependency relationships between inflective groups — lexical units within the subsets of words in a sentence. That is, in contrast to the system of Link Grammar Parser, which uses a dictionary containing the specifications describing relationships, in this case the link grammar is derived from the statistics.

The Turkish link parser considered in [13] is "not a lexical analyzer" in fact. At the first stage, a morphological analyzer is applied and some morphological descriptions are compared to the initial words. These descriptions are based on the analysis of the suffixes of words, which is natural for agglutinative languages. There are lexical items of only certain functionally important words. Then the links are established between morphological descriptions, not between the initial words. Apparently, it is possible to return to the initial sentence and carry the derived links to the words, but it is not considered in the work. This approach is used to describe the Turkish grammar in the terms of Link, but it is clear that it is applicable to other Turkic languages.

Finally, let's make a few remarks about the experiments of the authors of this paper. The link-grammar-4.7.12 developed at Carnegie-Mellon University [14] was taken as a basis of this work. It is an open multi-platform system. After some corrections and compilation in Visual Studio 10, we obtained the executable file of a program that can work with four languages: English, German, Russian and Lithuanian, though there are some drawbacks in its work, mainly related to the encoding.

English and German dictionaries are used for further development. A dictionary is replaced (for example, Kazakh and Turkish) by automatic translators. The specifications describing the links are manually entered into the dictionary, or available specifications are manually rectified. To work with dictionaries, the text editor Emurasoft EmEditor Professional 10.0.6 was used.

The question we have to answer is how many links should be used or what level of detail is necessary. For example, the English version has a separate link that connects the pronoun "he", "she" or "it" with a verb. It is known that in this case the verb must end with "s". Accordingly, the German version has a separate link connecting "du" (you) and a verb. The verb in this case must end with "st".

For Turkic languages, taking into account that they belong to the class of agglutinative languages, we find ourselves in a very difficult situation if we consider them in this level of detail. For the automatic translation, perhaps, it makes sense to develop such "heavy" analyzers, but for the information retrieval systems we can use a small limited set of links, such as proposed in [11].

7. CONCLUSION

The basic algorithm was tested in the iNetSearch system [6–8]. Ten simple queries from the field of inorganic chemistry have been generated. For each query, the lists of addresses with their description, usually returned to the user by a search system, have been loaded. On the basis of these short snippets, the resource estimation has been made. Statistics for comparison with a search engine (namely, http://www.nigma.ru, because it redirects the requests to other systems) have been obtained.

As a result of testing, on the average, the system allocated 5-15 qualitative relevant references out of 100 references received from Nigma.ru, accepted about 5 incorrect references as relevant and rejected others as irrelevant, which corresponds to reality. This demonstrates that the system can make filtration at a good level.

Then, two methods for the natural language constructions have been compared – the basic, used in the initial version of the iNetSearch system, and a new one, which takes into account the sentence rephrasing. The sources of the queries are as follows: a collection of scientific papers on more than 20 subjects and a collection of educational texts. Three different numerical characteristics were

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

considered to assess the quality of the search engine:

$$1)\ Precision = \frac{|Relevant \cap Retrieved|}{|Retrieved|}$$

$$2)\ Recall = \frac{|Relevant \cap Retrieved|}{|Relevant|}$$

$$3)\ Fall-out = \frac{|NotRelevant \cap Retrieved|}{|NotRelevant|}$$

The following notations are used:

Relevant is a set of documents from a collection relevant to the query;

NotRelevant is a set of documents irrelevant to the query;

Retrieved is a set of documents approved by the system.

On the average, the search system approves less irrelevant and more relevant documents. On the other hand, the method that uses rephrasing allowed us to improve the results of the iNetSearch system, but testing showed that this improvement is insignificant in comparison with the basic algorithm. Logical methods described in this paper are parts of further study, but they were not tested in detail in practice.

Let's make a few remarks about the limits of applicability of the methods. It is obvious that the proposed methods are applicable only to the sentences that can be quite correctly parsed by Link Grammar Parser. In other words, the methods are based on the assumption that the input of the system is a graph showing correct relations between entities. Note that Link Grammar Parser does not always generate an adequate diagram of links. Moreover, in most cases it puts out some diagrams, each of which is correct and therefore cannot be discarded. Most often it is due to the part-of-speech homonymy in the sentence or to the possibility of linking the words in different ways producing different interpretations of the sentence every time.

We can often interpret the obtained syntactic construction unambiguously and, according to our knowledge and experience, select the most appropriate diagram suggested by Link Grammar Parser. However, such knowledge is not built into the automatic interpreter, so it may put out more than one diagram for a given sentence, and it is impossible to know the number of the "correct" diagram in advance.

The suggested methods cannot match rephrased sentences if they contain formally different concept systems or the concepts are related to different semantic-syntactical relations, although the sentences may have the same meaning. In these cases, additional knowledge of the semantics has to be used, for example, the relative knowledge bases.

The studies in the Turkic languages stem from the need to analyze information in social networks, such as socio-economic, political, and radical Islamism. Investigations of this kind allow us to use Internet and social networks as a tool for influencing public sentiment and identifying social risks.

REFRENCES:

[1] G. Salton, *Automatic Information Organization and Retrieval*, 1968.

[2] G.V. Lezin, V.A. Tuzov, "The semantic analysis of the text in Russian: semantico-syntactical model of the sentence", *Economic-mathematical researches: mathematical models and information technologies*, Saint Petersburg: Nauka, no. 3, 2003, pp. 282-303.

[3] T.V. Batura, F.A. Murzin, "The machine-oriented logic methods of representation of semantics of the text in natural language", Novosibirsk: Publishing Company of NGTU, 2008.

[4] D. Temperley, D. Sleator, J. Lafferty, "Link Grammar Documentation", 1998, URL: http://www.link.cs.cmu.edu/link/dict/index.html (accessed on November 15, 2012).

[5] D. Sleator, D. Temperley, "Parsing English with a Link Grammar", Pittsburgh: School of Computer Science Carnegie Mellon University, 1991.

[6] F. Murzin, A. Perfliev and T. Shmanina, "Methods of syntactic analysis and comparison of constructions of a natural language oriented to use in search systems", *Bull. Nov. Comp. Center, Comp. Science*, no. 31, 2010, pp. 91-109.

[7] F. Murzin, A. Perfliev and T. Shmanina, "Methods of syntactic analysis and comparison of constructions of a natural language oriented to use in search systems", *Vestnik of Novosibirsk State Univ.*, vol. 9, no. 4, 2012, pp. 13-28.

[8] T.V. Batura, F.A. Murzin, A.A. Perfliev, T.V. Shmanina, "Methods of the increase of the efficiency of information search on the basis of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

syntactic analysis", Novosibirsk: Publishing Company of SB RAS, 2014.

[9] G.S. Lbov, "Methods of processing of polytypic experimental data", Novosibirsk: Nauka, 1981.

[10] A.A. Vikentiev, R.A. Vikentiev, "On the metrics for formulas containing polytypic variables and measures of denyty", *Proc. of the Second International Youth School-Conference "Theory and numerical methods of the decision of inverse and incorrect problems", Part 1,* 2011, pp. 192-209. URL: http://semr.math.nsc.ru/v8/c182-410.pdf

[11] U.A. Tukeyev, A.K. Melby, Zh.M. Zhumanov, "Models and algorithms of translation of the Kazakh language sentences into English language with use of link grammar and the statistical approach", *Proc. of IV Congress of the Turkic World Math. Society*, Baku, 1-3 July, 2011, p. 474.

[12] G. Eryiğit, K. Oflazer, "Statistical Dependency Parsing of Turkish", *Proc. of EACL 2006, 11th Conf. of the European Chapter of the Association for Comp. Linguistics*, Trento, Italy, 2006, pp. 89-96.

[13] O. Istek I. Cicekli, "A Link Grammar for an Agglutinative Language", *Proc. of Recent Advances in Natural Language Processing (RANLP 2007)*, Borovets, Bulgaria, 2007, pp. 285-290.

[14] Index of /downloads/link-grammar/4.7.12. [Electronic resource]. – Mode of access: http://www.abisource.com/downloads/link-grammar/4.7.12/ (accessed: 7 December 2014)

[15] I.M. Zhilin, *Synonymy in the syntax of the modern German language*, 1974.

[16] G.G. Pocheptsov, *Constructive analysis of a sentence structure*, 1971.

[17] B.P. Sukhotin, *Syntactic synonymy in modern Russian language*, 1960.