EURASIAN JOURNAL OF MATHEMATICAL AND COMPUTER APPLICATIONS
ISSN 2306–6172
Volume 4 , Issue 2 (2016) 14 – 22

USING THE LINK GRAMMAR PARSER IN THE STUDY OF TURKIC LANGUAGES

T.V. Batura, F.A. Murzin, D.F. Semich, S.K. Sagnayeva
S.Zh. Tazhibayeva, M.N. Bakiyev, A.S. Yerimbetova, A.M. Bakiyeva

Abstract Growing amount of information on the Internet and rapid development of social networks make the task of text processing increasingly actual. In this paper we propose an algorithm for the comparison of sentences and introduce certain measures of the closeness (similarity) between the sentences. The estimation of the relevance of documents should be based on the context of a search query and should not be limited only by keywords, their similarity or frequency. So proposed measures take into account lexical, syntactic and semantic relations between words. 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.

Key words: natural language processing, syntactic analysis, Link Grammar Parser, relevance, Turkic languages

AMS Mathematics Subject Classification: 68T50, 68P20, 68Q42

1 Introduction

Due to the fact of the increasing volumes of information networks the problem of improving the quality of the automatic information extraction becomes more and more topical. Many researchers [1, 2] introduce deep semantic analysis of texts for making the semantic images of texts, that can be the basis for document ranking. This approach, undoubtedly, is the most reasonable; however, it requires a careful and long-term work on the creation of suitable tools for natural language processing [3]. Therefore, a search for partial solutions, one of which is presented in this paper, is also useful.

Our main goal is to construct an algorithm for estimation the relevance of documents on the basis of sentences structure analysis. The relations between words built by Link Grammar Parser can be used to solve this problem [4, 5]. The algorithm for calculating the degree of similarity 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. The results of testing showed that the proposed algorithm efficiently solves the problem of information retrieval in English language texts. This is the main reason for our selection preference of the Link Grammar Parser. The Link Grammar Parser is a syntactic parser, based on link grammar [9]. The system assigns a syntactic structure for a sentence, which consists of a set of labeled links

connecting pairs of words. The main idea of link grammar allows us to work with an original theory of syntax and morphology at the same time. At the moment, there are plug-in dictionaries for English, Russian, Persian, Arabic, German, Lithuanian, Vietnamese, Indonesian language.

In this way, the Link Grammar Parser has a number of advantages: high speed of parsing, relatively easily customization for other languages, the ability of using morphological, syntactic and semantic levels of analysis simultaneously. One of the unsolved problems is the development of a parser like Link Grammar Parser for Turkic languages, which are most frequently used in the Internet, such as Kazakh and Turkish.

2 Parsers for Turkic languages

Many morphologic and syntactic parsers are developed nowadays. In particular, some approaches applied to agglutinative languages are described in the works [10-13].

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. [10]. Link Grammar plays an important role in the algorithm they are 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 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 [10]. 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 [11], 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 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.

The research [12] shows that the morphological and lexical information can improve parsing accuracy substantially. The proposed IG-based (inflectional group) models consistently outperform word-based models. This result has been obtained both for the probabilistic and the classifier-based parser, although the probabilistic parser requires careful manual selection of relevant features to counter the effect of data sparseness. A similar result was obtained in respect of lexicalized authors, in this case, although the improvement was only demonstrated in the classifier, which is probably due to its greater resistance was the scarcity of data based on the parsing. By combining a deterministic classifier based parsing approach with an adequate use of the model IG (inflectional group) on the basis of representations of morphological information and lexicalization, the authors concluded that they managed to achieve the highest accuracy for parsing the Turkish Treebank.

That is, in contrast to the system of Link Grammar Parser which uses a dictio-

nary containing the specifications that describe the relationship, 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. It should be noted that this kind of research was carried out by other authors [14-16].

The development of dependency parser based on the Kazakh language treebank is described in the paper [14, 15]. The most difficult step in this work is the creation of a free open-source dependency treebank. The authors note that their work is in the initial stage, so it is too early to talk about results.

Creating a treebank for the Kazakh language is a very laborious and time-consuming process that requires the work of a large number of linguists. Currently, we did not aim to create the Kazakh treebank. However, we do not exclude that this work will be done in the future.

Chagri Gultekin [16] initiates a new experience for morphologic segmentation, stemming, lemmatization, unknown words differentiation conversion of grapheme to phoneme, hyphenation and a morphological disambiguation on the data of the Turkish Language. The tools, which are promoted in the research, give new possibilities to build in a free open-source morphologic analyzer for Turkish natural language processing.

Of course, there is no doubt of the significance of Chagri Gultekin's new results on morphologic segmentation. However, there is a question of scholarly interest, whether it is possible to apply a set of open-source tools instruments for other Turkic languages of Oghuz, Karluk, Kypchak groups? In this respect it is interesting to compare the effectiveness of our research on a broad range of Turkic languages.

3 The Basic Algorithm of the Comparison of Sentences

We consider sentences as vectors, i.e. $\overline{x}=<x_1,..., x_n>$, $\overline{y}=<y_1,..., y_m>$, etc., where are words of each sentence. We suppose that they have been analyzed with the help of the Link Grammar Parser. 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 function words: articles, conjunctions, particles, interjections, etc. Let's assume now that I is a set of the pairs mentioned above, and its cardinality is $|I|=n$.

Then, let n_1, n_2 be the numbers of the links obtained as the result of the analysis of the sentences , respectively. As a measure of similarity of two sentences, it is possible to introduce

$$\mu_0(\overline{x},\overline{y})=\mathrm{n}/\max(n_1,n_2) \text{ or } \mu_1(\overline{x},\overline{y})=2\mathrm{n}/(n_1+n_2) \quad (1)$$

Thus, the method described above allows us to introduce measures of the similarity between sentences $\overline{x}$, $\overline{y}$. Note that both syntactic and semantic links can be used for the assessment of the similarity of the sentences. The description of these types of links is in the next two sections.

We discovered 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.

4 Links indicating syntactic features of words

The question we have to answer is how many links should be used and what level of detailing is suitable for us. 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 have "s" markers. Accordingly, the German version has a separate link connecting "du" (you) and a verb. The verb in this case must end with "st".

As for Turkic languages, they belong to the typological group of agglutinative languages. We come across the difficult problems with respect to the level of details. It makes a sense to develop such "heavy" analyzers for the automatic translation, however for the information retrieval systems a small limited set of links can be used, such as proposed in [10]. We have identified the following basic connections in the Kazakh and Turkish languages: AS is an attribute of a subject; AO is an attribute of an object; E is an adverbial modifier; J connects a postposition and a noun; OV is a direct object; OJV is an indirect object; S connects a subject and a predicate.

If we consider syntactic features of words in a sentence, then each part of speech can be associated with a formula of possible connectors: a noun may act as a subject connected to an attribute; a verb has to be at the final position and end a sentence, etc. Here is an example of a sentence structure in the Turkish language: <N_S>: {AS-} & {OV+} & S+. Besides, a noun may act as an object, on the left of which is an attribute, on the right is a postposition and predicate. Such structure is generally described by the formula: <N_O>: {AO-} & {OV+} & {OJV+}.

Let us consider the following sentence

Адамдар алма жеді.

Адам-дар алма же-ді.

People-PL apple eat-V-PST

People an apple ate.

The parser identifies two syntactic (S3p, OV) and two morphological (Np, Va3p) links. An example of this parsing is shown in Picture 1.

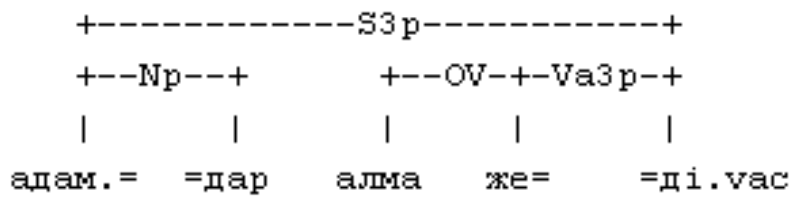

Figure 1. Let us consider the following sentence

5 Model of semantic markup of sentences

In order that to make a transition from the morphological and syntactic to semantic links, it is more convenient to carry out further considerations in terms of predicates. We have two-place predicates because we use link grammar. Thus syntactic links mentioned above, in some cases, can be considered in the form of predicates: AS (adjective, noun); AO (adjective, noun); E (adverb, verb); OJV (Nd (noun) | Na (noun) | Ni (noun) | Nl (noun) | Nb (noun), verb); S (Nn | Pn), verb), etc. Note that under this approach the predicates OV (x, y) and OJV (x, y) contain information about the verbal coordination, i.e. they depend on the use of a specific case before the certain verb. In the future, we plan to carry out additional study of verbal coordination in the Kazakh and Turkish languages. Now it is possible to consider the semantic predicate of possession: OF (Possessor, Possessed) = OF (Ng (noun) | Pg (pronoun), Np3 (noun)). The predicate OF (x, y) describes, for example, the phrase: kadının elbisesi ("women's dress", i.e. dress which belongs to the woman), where kadın is a stem of a word ("female"); ın is a genitive suffix; elbise is a stem of a word ("dress"); si is a possessive suffix. Consider the sentence: Ben kardeşin kitabını okuyorum. (I am reading the brother's book.). Let us write this sentence with the help of the predicates: READ (ben, OF (kardeşin, kitabını)). The predicate OF enables emphasize the possessive pronouns. Picture 2 shows a parsing example, Менің қарным ашқан жоқ. (I am not hungry.), containing the first person possessive pronoun (the link OF1 is responsible) and the negative form of the verb (the link VN is responsible).

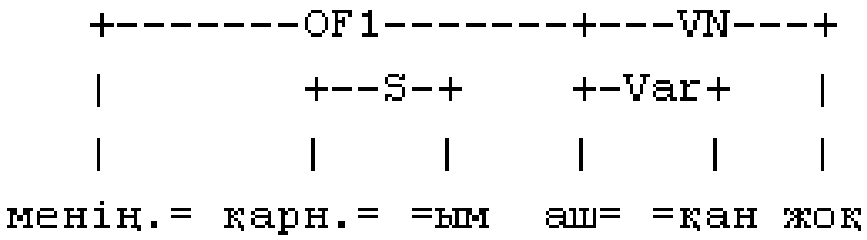

Figure 2. Possessive pronouns in Kazakh language

A sentence parsing example with the possessive pronoun Senin ne istedigini bilmiyorum. (I don't know what you want.) is shown below.

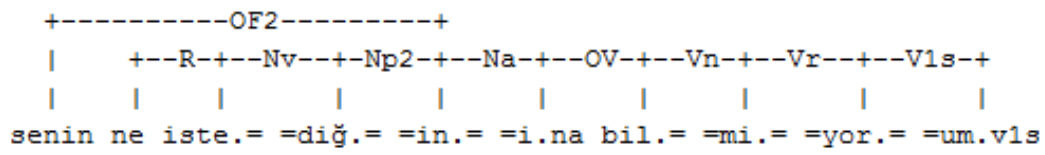

Figure 3. Possessive pronouns in Turkish language

Semantic predicates of place LOC (verb, adverb) and time of action TIME (verb, adverb) are interesting from the perspective of further research. The predicate FOR (Ng (noun) | Pg (pronoun), postposition) describes a combination of a postposition "için" with a noun or pronoun in the genitive case.

6 Conclusion

The study of the Turkic languages is stipulated by the necessity to analyze information from social networks: socio-economic, political, about radical Islamism, etc. Investigations of this kind allow us to use Internet and social networks as a tool for influencing public sentiment and identifying social risks.

In this paper, we propose an algorithm for the comparison of texts (as a sequence of sentences) and estimation of their similarity. This method is applicable only to the sentences that can be quite correctly parsed by the Link Grammar Parser. The proposed measure takes into account lexical, syntactic and some semantic relations between words.

We carried out experiments to assess the relevance of texts in Turkish and Kazakh languages to the search query. The volume of the Kazakh dictionary for the Link at the moment is rather small, about 500 words and 100 affixes. The size of texts in both languages in our experiment was 11-27 Kb. For example, the following results of the comparison of the Kazakh texts were obtained:

(text_1,text_2)=0.4727; (text_1,text_3)=0.4364; (text_1,text_4)=0.4;
(text_2,text_3)=0.766; (text_2,text_4)=0.2215; (text_3,text_4)=0.2123.

The results of the comparison of the Turkish texts are as follows:

(text_1,text_2)=0.6041; (text_1,text_3)=0.5833; (text_1,text_4)=0.75;
(text_2,text_3)=0.1305; (text_2,text_4)=0.1188; (text_3,text_4)=0.2025.

In the current time is until not easy that thresholds expediently to use for separation of relevant and not relevant texts. Of course, thresholds may depend on themes texts and many other factors. Also it is possible to apply various algorithms of machine learning to define thresholds.

The accuracy of the Link Grammar Parser is mostly dependent on the completeness of dictionaries. In the preparation of dictionaries it is necessary to take into account some specific morphological characteristics with respect the agglutinative structure of the Turkic languages.

During the research, we also faced with problems of linguistic ambiguity: how to describe the homogeneity of the sentence part, using the link grammar, how to cope with homonymy of stems and affixes, etc. And such problems we still have to solve.

References

[1] Salton G., *Automatic Information Organization and Retrieval*, 1968, 514.

[2] Lezin G. V., Tuzov, V. A. , *The semantic analysis of the text in Russian: semantico-syntactical model of the sentence, Economic-mathematical researches: mathematical models and information technologies*, Is. 3, Nauka, Saint Petersburg (2003) 282–303. (in Russian)

[3] Batura T. V., Murzin F. A., *The machine-oriented logic methods of representation of semantics of the text in natural language*, A.P. Ershov Institute of Informatics Systems SB RAS, Publishing Company of NGTU, Novosibirsk. (2008) 248. (in Russian)

[4] Temperley D., Sleator, D., Lafferty J., *Link Grammar Documentation*, (http://www.link.cs.cmu.edu/link/dict/index.html)

[5] Sleator D., Temperley D., *Parsing English with a Link Grammar*, School of Computer Science Carnegie Mellon University, Pittsburgh, (1991) , 93.

[6] Murzin F., Perfliev A., Shmanina T., *Methods of syntactic analysis and comparison of constructions of a natural language oriented to use in search systems*, Bull. Nov. Comp. Center, Comp. Science, Iss. 31 (2010), 91-109.

[7] Murzin F., Perfliev A., Shmanina T., *Methods of syntactic analysis and comparison of constructions of a natural language oriented to use in search systems*, Vestnik of Novosibirsk State Univ. Ser.:Information Technologies, Novosibirsk, vol. 9, Iss. 4. (2012), 13-28. (in Russian)

[8] Batura T. V., Murzin F. A., Perfliev A. A., Shmanina T. V., *Methods of the increase of the efficiency of information search on the basis of syntactic analysis*, A.P. Ershov Institute of Informatics Systems SB RAS.: Publishing Company of SB RAS, Novosibirsk. (2014), 76. (in Russian)

[9] Temperley D., *An Introduction to the Link Grammar Parser*, (2014), (http://www.abisource.com/projects/link-grammar/dict/introduction.html)

[10] Tukeyev U. A., Melby A. K., Zhumanov Zh. M., *Models and algorithms of translation of the Kazakh language sentences into English language with use of link grammar and the statistical approach*, Proceedings of IV Congress of the Turkic World Math. Society, Baku. (2011), 474.

[11] Gülşen Eryiğit, Kemal Oflazer, *Statistical Dependency Parsing of Turkish*, Proceedings of EACL 2006, 11th Conf. of the European Chapter of the Association for Comp. Linguistics, Trento, Italy (2006), 89-96.

[12] Gülşen Eryiğit, Joakim Nivre, Kemal Oflazer, *Dependency Parsing of Turkish*, Computational Linguistics, vol. 34, N. 3 (2008), 357–389.

[13] Ozlem Istek, Ilyas Cicekli, *A Link Grammar for an Agglutinative Language*, Proceedings of Recent Advances in Natural Language Processing (RANLP 2007), Borovets, Bulgaria. (2007), 285-290.

[14] Tyers F. M. and Washington J., *Towards a free/open-source universal-dependency treebank for Kazakh*, Proceedings of the 3rd Conference on Turkic Languages (TurkLang 2015), (2015), 276-289.

[15] Washington J. N., Salimzyanov I. and Tyers F. M., *Finite-state morphological transducers for three Kypchak languages*, Proceedings of the 9th Conference on Language Resources and Evaluation, LREC2014. (2014), 3378-3385.

[16] Çağrı Çöltekin, *A Set of Open Source Tools for Turkish Natural Language Processing*, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) Ed. by N. Calzolari et al. (2014), 1079-1086.

T.V. Batura,
A.P. Ershov Institute of Informatics Systems, Russian Academy of Sciences, Siberian Branch,
Novosibirsk State University,
6, Acad. Lavrentjev pr., Novosibirsk 630090, Russia,
Email: tatiana.v.batura@gmail.com,

F.A. Murzin,
A.P. Ershov Institute of Informatics Systems, Russian Academy of Sciences, Siberian Branch,
Novosibirsk State University,
6, Acad. Lavrentjev pr., Novosibirsk 630090, Russia,
Email: murzin@iis.nsk.su,

D.F. Semich,
A.P. Ershov Institute of Informatics Systems, Russian Academy of Sciences, Siberian Branch,
6, Acad. Lavrentjev pr., Novosibirsk 630090, Russia,
Email: deiman32@ngs.ru,

S.K. Sagnayeva,
L.N. Gumilyov Eurasian National University,
2, Satpayev St., Astana 010008, Kazakhstan,
Email: sagnaeva_tar@mail.ru,

S.Zh. Tazhibayeva,
L.N. Gumilyov Eurasian National University,
2, Satpayev St., Astana 010008, Kazakhstan,
Email: tazibaeva_szh@enu.kz,

M.N. Bakiyev,
L.N. Gumilyov Eurasian National University,
2, Satpayev St., Astana 010008, Kazakhstan,
Email: murat26261957@mail.ru,

A.S. Yerimbetova,
L.N. Gumilyov Eurasian National University,
2, Satpayev St., Astana 010008, Kazakhstan,
Email: aigerian@mail.ru,

A.M. Bakiyeva,
Novosibirsk State University,

2, Pirogova St., Novosibirsk 630090, Russia,
Email: m_aigerim0707@mail.ru

Received 08.06.2016, Accepted 20.06.2016