Uluslararası
Uygur Araştırmaları Dergisi
2017/10

COMPUTATIONAL LINGUISTICS AND ADAPTATION OF TURKIC LANGUAGES TO COMPUTER

Bilişimsel Dilbilimi ve Türk Dillerinin Bilgisayara Uyarlanması

Murat ORHUN[**]

Abstract
This article describes computational linguistics briefly, and explains Turkic language studies in this field using Uyghur language as an example. With developing computer technologies, many software has been implemented in order to complete some tasks in place of human. For example, translate from one language to another, or translate from one language to more than one languages at the same time, correcting or editing texts, analyzing documents, converting speeches into texts or converting texts into speeches etc. Until now, there are many successful researches have been done on different languages such as English, Japanese, Arabic, Turkish, Chinese, French and Russian etc. In Turkic languages, especially in Turkey Turkish, though there are some important researches have been done, other Turkic languages still at a beginning stage. Though, Uyghur language belongs to Turkic language family and it has common properties with other languages, however research results about other Turkic languages cannot be applied to Uyghur language directly. As a natural language, Uyghur language has many special properties those (are) different from other Turkic languages. This paper summarizes some computer based researches about Uyghur language and use them as a part of general machine translation system of the Turkic world.
Keywords*: Turkic Languages, Machine Translation, Turkic World, Morphological analyzer, Turkic Corpus.*

Özet
Bu makalede bilişimsel dilbilimi kısaca anlatılmıştır ve bu alanda Türki diller ile ilgili yapılan çalışmalar Uygur dili örnek verilerek açıklanmıştır. Teknolojinin ilerlemesi ile insanların yapması gereken işlevleri yerine getirecek pek çok yazılımlar geliştirilmiştir. Örneğin bir dilden başka dile aktarma, ya da bir dilden aynı anda birden fazla dile aktarma, metin düzeltmek ya da biçimlendirmek, dosya analiz etmek, sesi yazıya dönüştürmek ya da yazıyı sese dönüştürmek gibi çalışmalar başarılı bir şekilde yazılımlar tarafından gerçekleştirilmektedir. Şimdiye kadar İngilizce, Japonca, Arapça, Çince, Fransızca ve Rusça gibi diller ile ilgili pek çok araştırmalar yapılmıştır ve başarılı sonuçlar elde edilmiştir. Türki diller içinde, özellikle Türkiye Türkçesi ile ilgili bazı önemli araştırmalar yapılsa da, diğer Türki diller ile ilgili araştırmalar henüz başlangıç aşamasındadır. Gerçi Uygurca Türki diller ailesine ait ve diğer Türki diller ile ortak özelliklere sahip olsa da, diğer Türki diller ile ilgili yapılan araştırma sonuçları doğrudan Uygurcaya uygulanamaz. Doğal dil yönünden, Uygurca diğer dillerden farklı ve azımsamayacak kadar özellikler bulunduruyor. Bu makalede, Uygurca ile ilgili yapılan bilgisayara dayalı araştırmalar özetlenmiştir ve bu çalışmaların Türk diller ile ilgili genel çeviri sistemlerinde kullanılması hakkında öneri de verilmiştir.

Anahtar Sözcükler: *Türk dilleri, Bilgisayarlı Çeviri, Türk Dünyası, Biçimbilimsel çözümleyici, Türk Derlemi.*

**Asst. Prof. Dr., Computer Engineering Dept., Istanbul Bilgi University, Istanbul-TURKEY, E-posta: murat.orhun@bilgi.edu.tr.*

Introduction

In a natural language, there are many different ways to explain an idea and there are hundreds of languages in the world today. In some countries, more than one languages have been accepted as official languages. For example, English and French must be used in all official documents in Canada. Therefore, definitively it is necessary to translate all documents into both of these languages. While speaking in a meeting or in a conference, it is necessary to translate one sentence into other language immediately. And if it is required to translate a longer article naturally, it (will) take long time to get the translated document. At the same time the quality of the translation also important. Because of these reasons, translations have been very important task from early ages. Scientists have been trying to find a general solution to translations for a long time. With the inventions of computer, language translation and their research have become one of the hot topic in science and it is called computational linguistics. Both of scientists and linguists tried to translate one language into another with a computer program. The first computer based translation system had been implemented from Russian into English about 60 sentences (Chéragui 2012:161-163). The translation result was a great success and scientists thought it was possible to implement a machine translation system for general purpose in 3-5 years. After the real project had been started, the progress of the project was too slow and couldn't get the expected results after 10 years of research. Because of this reason, the famous ALPAC report was issued (Hutchins 1995:439-440). With this report USA government had reduced the fund for researching computer based translation or machine translation. Beginning in the late 1980s, computer technologies developed better than 1960s and computers with large memory and high speeds are available for less money. Hence, computer based translation or computational linguistics has been become one of the hottest topic of the contemporary science. Machine Translation (MT) is a sub-filed of Natural Language Processing (NLP) and NLP is a field of Artificial Intelligence. The purpose of the machine translation is to translate one natural language into another natural language with a software (computer programs) without any help of human. Unfortunately, it is too difficult to implement such a translation system. The main reason is, it is a natural language first and there are hundreds of cases or shapes to explain an idea. For example, idioms, humors, phrases and poems etc. Meanwhile some explanations have been related to cultural and habitual activities or even speaking tones. At the moment it is not possible to implement such a Fully Automatic High Quality (FAHQ) translation system. Though Fully Automatic High Quality translation system is (not) possible, some special machine translation systems have been implemented and have been used actively in every life and researches. For example, the Météo, it is a machine translation system was developed for the translation of weather bulletins from English to French issued by the meteorological institutes in Canada (Chandiox 1976:127-129), the English-Japanese machine translation system of the titles of scientific and engineering papers (Nagao and Tsujii et al., 1982:245-246) etc. Apart from these, some machine translations systems have been implemented for general purpose even don't meet the FAHQ criteria. For example, the RUSLAN system (Hajič 1987:113-117), that translates from Czech language to Russian and the CESILKO system (Hajič and Hric et al., 2000:7-12), which translates from Czech language to Slovak language. The translation quality is different according to closeness of different languages. For example, the translation corrects of the RUSLAN system is about 40 percent while the CESILKO system's is about 90 percent.

This paper supposed to give some information of computer based translation that related to Uyghur language in order to explain recent development of computational research of Turkic languages. The rest of the paper is organized as follows. The next section gives some summarization of machine translation that related to Turkic and Uyghur languages. Section three introduces and discusses some problems related to machine translation about Uyghur

language. Section four gives a brief conclusion about computational linguistics and their effects on Turkic studies.

Related Works

Uyghur language is a Turkic language and it belongs to the Ural-Altaic language family. Almost all Turkic languages have the same grammatical structure except simple difference. One of the main differences between these languages is about new words that those accepted from other languages. Russian words appear in Central Asian Turkic languages while Chinese, Arabic and Persian words appear in Uyghur language and English words appear in Turkey Turkish etc. Because of these reason, some differences appear when adding suffix or prefixes to a word. In natural language studies, it is the first step to analyze a word correctly with its morphemes. Turkic languages are agglutinative and heavily inflected language. It means a word could take no limited suffixes theoretically and changes some characters in order to harmonize vowel and constants. For example:

OSMANLILAŞTIRAMAYABİLECEKLERİMİZDENMİŞSİNİZCESİNE

This word can be broken into morphemes as follows:

OSMAN+LI+LAŞ+TIR+AMA+YABİL+ECEK+LER+İMİZ+DEN+MİŞ+SİNİZ+CESİNE

This is a famous example in Turkish rather exaggerated and it means “as if you were of those whom we might consider not converting into an Ottoman” (Oflazer 1995:1). The root of this word is “OSMAN” and the rest of the words are suffixes. Such examples could be found out in other Turkic languages as well. In order to work on a word, it is necessary to understand that word correctly and there are millions of different combinations of words with its possible prefixes and suffixes. While attaching a suffix to a word, that word's or phrases` category will be changed according type of a suffix. Therefore, these changes will affect the structure of a whole sentence (Oflazer 1995:1-2). In Turkic language family, Turkey Turkish is one of the most studied language in computer science. Turkish language was the first language that its morphology had been analyzed with a computer. Because all Turkic languages belong to the same language family, also they are very closely related to each other, some technical researches could be applied to other Turkic language with some modifications. For example, Turkmen (Tantuğ and Adalı et al., 2006a), Crimean Tatar (Altıntaş and Çicekli 2001), Uyghur (Orhun and Tantuğ et al. 2009a, Orhun and Tantuğ et al. 2009b), Kazakh (Kessikbayeva and Çiçekli 2014) and Qazan Tatar (Gökgöz and Kurt et al., 2011) language morphological analyzers have been implemented based on the Turkish morphological analyzer (Oflazer 1995). Turkic languages are agglutinative language; therefor usually more than one solutions are generated when a word is analyzed. Because of this reason, morphological ambiguity will appear to decide which solution is correct (Oflazer etc. 1996). For example, the Uyghur word “yazmaqchi” (will write) will be generated by following solutions when analyzed with the Uyghur morphological analyzer.

yazmaqci: yaz+Verb+Pos+Fut+A3sg
yazmaqci: yaz+Verb+Pos^DB+Noun+Inf1+A3sg+Pnon+Nom^DB
+Adj+Agt
yazmaqci: yaz+Verb+Pos^DB+Noun+Inf1+A3sg+Pnon+Nom^DB
+Noun+Agt+A3sg+Pnon+Nomyazmaqci:yaz+Verb+Pos
+Inten+A3sg

The first solution explains, the root of the word is “yaz” (write), it is a verb, positive, future tense and in third person singular form. The second solution explains, the root is “word”, positive, with adding the “maq” suffix, the word has been become pronoun, also this

pronoun has been become adjective with the adding the “ci” suffixes. Rest of the solution could be analyzed with the same way. To solve such morphological disambiguation problems, there are some important researches have been done with the fund of government for Turkish language (Oflazer and Hakkani-Tür et al., 1996, Hakkani-Tür and Oflazer et al., 2002). After morphological analyzers have been implemented (for) both of the source and target languages and disambiguation problem has been solved, then a simple machine translation system could be implemented. For implementing a translation system, rule based or statistical method can be used, or a hybrid system can be used (Tantuğ 2007). Tantuğ (Tantuğ 2006b) has implemented a morphological analyzer for the Turkmen language first. Because there are some differences between Turkmen and Turkish sentence, some rules have been defined to correctly replace some Turkmen suffixes with Turkish suffixes. After such rules have been defined, Turkmen root words have been translated into Turkish root word. As a natural language, a word can be translated into another language more than one word. This case creates ambiguation problem. Before getting the correct translation, one of the best word should be selected according to sentence meaning. To solve this problem, rule based methods cannot provide good solution. Therefore, statistical methods have been suggested based on corpora. In order to decide the best word, computers calculate a words frequency that may appear a relatively similar sentence in the corpora and selects the word with the high frequency (see Fig.1).
In this Figure, the calculation process described to choose three words that translated from Turkmen to Turkish such as. “ne” or “kim”, “insan” or “adam”, “konuş” or “söyle”. Even Turkish and other Turkic languages are closely related to each other, the Turkish language research results can not be applied to them directly.

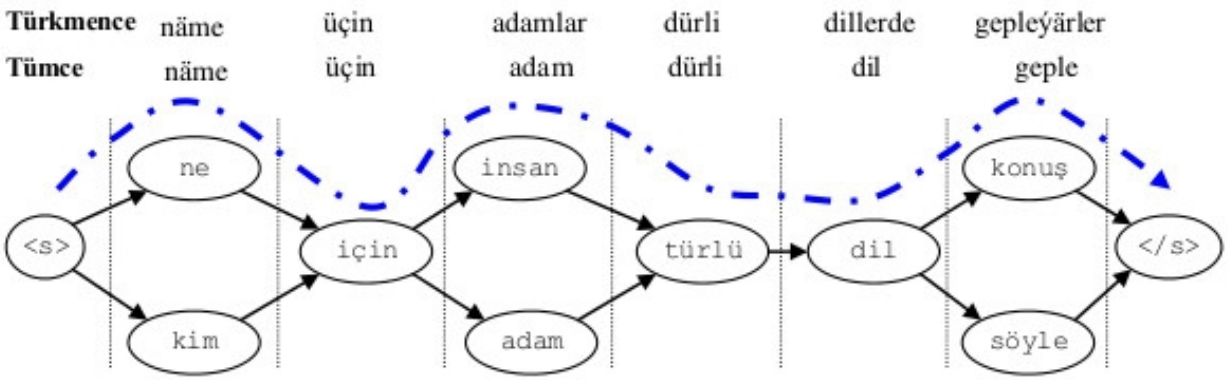

Fig.1 The process of decoding the most probable target language sentence (Tantuğ et al., 2006b)

Hence-forth, Uyghur language has been studied independently and some primary results have been achieved by a system that translates from Uyghur to Turkish (Orhun 2010). In this translations system, a rule based word sense disambiguation model implemented instead of statistical based methods. Therefore, the calculation speed is higher than statistical methods. For example, the sentence “men qelem aldim” means “I bought a pen” can be analyzed as follows:

men: men+Pron+Pers+A1sg+Pnon+Nom
bir: bir+Num+Card
qelem: qelem+Noun+A3sg+Pnon+Nom (pen)
qelem: qele+Noun+A3sg+P1sg+Nom (my castle)

aldim: aldi+Noun+A3sg+P1sg+Nom
aldim: al+Verb+Pos+Past+A1sg

In these solutions, the word “qelem” has produced two different solutions. To decide which one is correct, the firs word “men” should be analyzed. The word “qelem” related to subject of the sentence “men” and subject doesn't take any personal suffix “P1sg”. Therefore, the solution result “my castle” will be discarded and other solution considered as a correct one. The drawback is, only limited number of rules have been defined and cannot give correct result for ignored cases.

Restrictions of Machine Translation

To implement a machine translation system from Uyghur to Turkish, a rule based morphological analyzer has been implemented for nouns (Orhun et al., 2009a) and verbs (Orhun et al., 2009b). With this analyzer, contemporary Uyghur words have been analyzed about 88 percent correctly at this current version. The reason is that the correctness is not so high, there are some words that they don't belong to Uyghur language originally and they couldn't have analyzed with general rules. Not only there are some Chinese and English words appear in the contemporary Uyghur language, but also some Persian and Arabic words as well. Another reason is, The Uyghur verbs have very complex structure and there are a lot of auxiliary verbs as well. Whenever some suffixes have been attached to a verb, the root word or formed words will be inflected. Also, the auxiliary verbs have not been considered in (Orhun et al., 2009b). Therefore, it is still an open topic to be studied. Because the morphological problem has not been solved properly, it is difficult to solve the disambiguation problem. Without solving the disambiguation problem, it is not possible to get the correct translation of the source language. The system introduced in (Orhun 2010) includes some rules that defined based on classification of the words, which meaning based on morphological analyzes. For example, let us analyze the following Uyghur sentence (he/she will write a letter).

u: u+Pron+Pers+A3sg+Pnon+Nom
xet: xet+Noun+A3sg+Pnon+Nom
yazmaqci: yaz+Verb+Pos+Fut+A3sg

yazmaqci: yaz+Verb+Pos^DB+Noun+Inf1+A3sg+Pnon+Nom^DB
+Adj+Agt

yazmaqci: yaz+Verb+Pos^DB+Noun+Inf1+A3sg+Pnon+Nom^DB
+Noun+Agt+A3sg+Pnon+Nom

yazmaqci: yaz+Verb+Pos+Inten+A3sg

After that sentence has been analyzed, the word “yazmaqci” (going to write) will be analyzed with four different solutions. Actually, one solution that is in the “noun”, yaz (summer) form has been eliminated.

yaz: yaz+Noun+A3sg+Pnon+Nom

The reason is, whenever a word attached the suffix “maqci”, then that word is definitely a verb. Because the “maqci” suffix used (to) create a future tense from a verb. Therefor the solution with the noun property will be eliminated automatically even there is a possibility that it could be resolved as a noun. If a system suggested that works without a rule definition, then a statistical system must be necessary. Unfortunately, there is not some general corpora today for Uyghur language. Though some researches have been started

using corpus constructing, it is not available for public or academic research (Aibaidulla and Kim-Teng 2003). The task constructing a corpus is a very expensive task and it takes a long time. After morphological disambiguation has been solved, the translation of the source language will be searched in a bilingual target language and target sentence will be created. To improve machine translation system about Uyghur language some researches have been done recently. For example, analyze words according to their syllabic properties (Orhun 2016). There are six different syllabic forms which have been found out for Uyghur origin words while 5 different forms for adopted words (Orhun 2016). In some cases, some Uyghur words cannot be solved with simple classification. To classify or analyze them, it is important to find out those words roots or entomology (Abdulla 2016). As supposed in other languages, corpus based solution is suggested to Uyghur language in order to make statistical calculate and use more machine learning algorithms.

In general, all Turkic language have limited number of root words, and other words will be created with adding suffixes. Once root words have been translated correctly, then rest of the words could be accessed by applying different rules or some statistical calculation results. In order to get roots and solve disambiguation problem, it is necessary to have high quality full functional morphological analyzer. In machine translation, it is very critical task to convert contents of the source sentence semantically into target language. This process includes solving sentence structure of the target language. In most of successful machine translation systems, translation algorithm handles sentence structure tasks which is from both source and target language.

For Turkic languages, there is a Treebank corpus (Oflazer and Say et al., 2003), for the Turkish language only and this is a drawback for implement translation systems between different Turkic languages.

Conclusion

In this paper, some computer based language analyzing methods have been introduced briefly that related to Turkic languages with explaining recent researches about Uyghur language. As a result, there is not a full functional morphological analyzer for the Uyghur language. Therefore, it is still early to do large scale computational research on this language. Because of this reason it is not possible to (use) a well implemented machine translation system from or to Uyghur language at the moment. In this Internet age, machine translation is not avoidable tendency. For example, the Google translator is one of the practical example that used in everyday life. Even the Google translator cannot give correct or well-structured translation, still it gives brief information about source texts. All Turkic languages are close to each other and if a common corpus is implemented for all all Turkic languages, it is possible to implement a general machine translation system for all Turkic languages.

References

ABDULLA A. (2016). “Chaghatay Language is The Bridge Between Old Turkic and Modern Uyghur Language”. *International journal of Uyghur Studies*, S. 7. p: 1-7.

AIBAIDULLA Y. and Kim-Teng Lua (2003). "The development fo Tagged Uyghur Corpus". *Proceedings of PACLIC17*. Sentosa, Singapore, p:228-234.

ALTINTAŞ K., and Çiçekli, İ. (2001). “A Morphological Analyser for Crimean Tatar”. *in Proceedings of the 10th Turkish Symposium on Artificial Intelligence and Neural Networks*, TAINN North Cyprus, p: 180-189.

CHANDIOUX J. (1976*). “*Météo: Un système opérationnel pour la traduction automatique des bulle tins météorologiques destinés au grand public*”. Meta,* vol. 21, p: 127-133.

CHÉRAGUI, M. Amine (2012). "Theoretical Overview of Machine translation". Proceeding: *International Conference on Web and Information Technologies* (ICWIT), pp:160-169.

GÖKGÖZ E; Kurt A.; Kulamshaev K.; Kara M. (2011). "Two-Level Qazan Tatar Morphology". *1st International Conference on Foreign Language Teaching and Applied Linguistics*, May 5-7, Sarajevo, pp: 428-342.

HAJIČ J. (1987). "RUSLAN - An MT System Between Closely Related Languages". in *Third Conference of the European Chapter of the Association for Computational Linguistics* (EACL'87) Copenhagen, Denmark, pp: 113-117.

HAJIČ, J.; Hric J.; Kuboň V. (2000). "Machine translation of very close languages". in *Proceedings of the sixth conference on Applied natural language processing*, Morgan Kaufmann Publishers Inc, pp: 7-12.

HAKKANI-TÜR D.Z; Oflazer K.; Tür G., (2002). "Statistical Morphological Disambiguation for Agglutinative Languages". *Computers and the Humanities*, Vol.36, No.4, pp: 381- 410.

HUTCHINS J. (1995). *Machine Translation: A Brief History, Concise history of the language sciences: from the Sumerians to the cognitivists*. Edited by E.F.K., Koerner ve R.E.Asher, Oxford, Pergamon Press, pp: 431- 445.

KESSIKBAYEVA G.; Çiçekli İ. (2014). "Rule Based Morphological Analyzer of Kazakh Language". *Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM Baltimore*, Maryland, USA, pp: 46–54.

NAGAO M.; Tsujii J.; Yada K.; Kakimoto T. (1982). "An English Japanese Machine Translation System Of the Titles Of Scientific And Engineering Papers". *International Conference On Computational Linguistics*, Porceedings of the 9th conference on Computational Linguistics- Vol.1, Prague, Czechoslovakia, pp: 245- 225.

OFLAZER K.; Tür G., (1996). "Combining Hand-crafted Rules and Unsupervised Learning in Constraint-based Morphological Disambiguation". *Proceedings of the ACL-SIGDAT Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, USA, pp: 69-81.

OFLAZER K. (1995). "Two-level Description of Turkish Morphology", *Literary and Linguistic Computing*, Vol. 9, pp. 137-148.

OFLAZER K.; Say B.; Hakkin-Tür D.Z. and Tür G. (2003). Building a Turkish Treebank, V.20 of the series Text, Speech and Language Technology, pp: 261-277.

ORHUN M. (2010). "Machine Translation from Uyghur to Turkish". *PhD Thesis*, Istanbul Technical University. Istanbul, Turkey, pp:1-180.

ORHUN M. (2016), "Word Identification According to Syllabic Property". Southeast Europe Journal of Soft Computing, VOL.5, NO.2, September, pp: 11-15

ORHUN M.; Tantuğ A.C; Adalı E. (2009b). "Rule Based Tagging of the Uyghur Verbs". *Fourth International Conference on Intelligent Computing and Information Systems*. Faculty of Computer &Information Science, Ain Shams University, Cairo, Egypt, pp: 811-816.

ORHUN M.; Tantuğ A. C and Adalı E. (2009a). "Rule Based Analysis of the Uyghur Nouns". *International Journal of Assian Language Processing* 19(1), pp: 33-43.

TANTUĞ A. C.; Adalı E. and Oflazer K. (2006a). "Computer Analysis of The Turkmen Language Morphology ", *Proceedings of the 5th International Conference on Natural Language Processing*, FinTAL, Turku, Finland, Vol. 4139, pp: 186-193.

TANTUĞ A. Cüneyd; Adalı E. and Oflazer K. (2006b). “A Prototype Machine Translation System Between Turkmen and Turkish”. *Fifteenth Turkish Symposium on ArtificialIntelligence and Neural Networks*. TAINN Gökova. Muğla. Türkiye, pp:109-116.

TANTUG A. C. (2007). "A Hybrid Model For Machine Translation Between Agglutinative and Related Languages". *PhD Thesis*. Istanbul Technical University. Istanbul. Turkey. pp: 1-145.