"DILMANC" is the 1st MT System for Azerbaijani

Rauf Fatullayev
National E-Governance Project
Baku, Azerbaijan
fatullayev@gmail.com

Ali Abbasov
National Academy of Sciences
Baku, Azerbaijan
ali@dcacs.ab.az

Abulfat Fatullayev
Institute of Linguistics
Baku, Azerbaijan
fabo@box.az

1 Turkic languages

Turkic languages – modern Turkish, Azerbaijani, Kazakh, Uzbek, Kyrgyz, Turkmen, Tatar etc. – form one of the largest language groups and these languages are very close in morphological and syntactic levels. These languages have rich morphological structures (it is possible to form very large number of word-forms from a given stem by adding the rich set of simple and compound suffixes) and as a result it is very difficult to create an MT system for these languages (Fatullayev et al. 2008). Despite the morphological and syntactic closeness, the vocabulary differences are also characteristic for these languages and consequently the development of the MT systems for these languages is being carried out on two different directions:

1. Development of the MT systems among Turkic languages;
2. Development of the MT systems from/into Turkic languages into/from languages not belonging to this group.

2 Dilmanc MT system

Researches on the development of the practically useful NLP systems for Azerbaijani are being carried out since 2005 within the joint project of the Ministry of ICT of Azerbaijan and UNDP-Azerbaijan (Dilmanc project). The 1st version of the Azerbaijani MT system (Dilmanc MT system, www.dilmanc.az) is already developed and this MT system is one of the first softwares for Turkic languages. Dilmanc MT system for the present can translate on three directions – Azerbaijani-English, English-Azerbaijani and Turkish-Azerbaijani.

Most languages of Turkic group are still less investigated languages, except modern Turkish [Oflazer (http://people.sabanciuniv.edu/oflazer/pubs.html), Cicekli (http://www.cs.bilkent.edu.tr/~ilyas/pubs.html)]. But, despite many researches dedicated to different aspects of the development of the MT system from modern Turkish the whole technology permitting to automate the translation process from this language has not been developed yet. Technology for the automation of the translation process from Azerbaijani developed within Dilmanc project is the first in this field. Most of the necessary works (development of MT dictionaries, formal grammar for Azerbaijani, algorithms for the automation of the translation process from/into Azerbaijani, synthesizer and analyzer algorithms of the Azerbaijani sentences etc.) are carried out for the first time (Fatullayev, 2005; Abbasov and Fatullayev, 2007; Fatullayev et al. 2008).

Dilmanc is a hybrid MT system developed on the

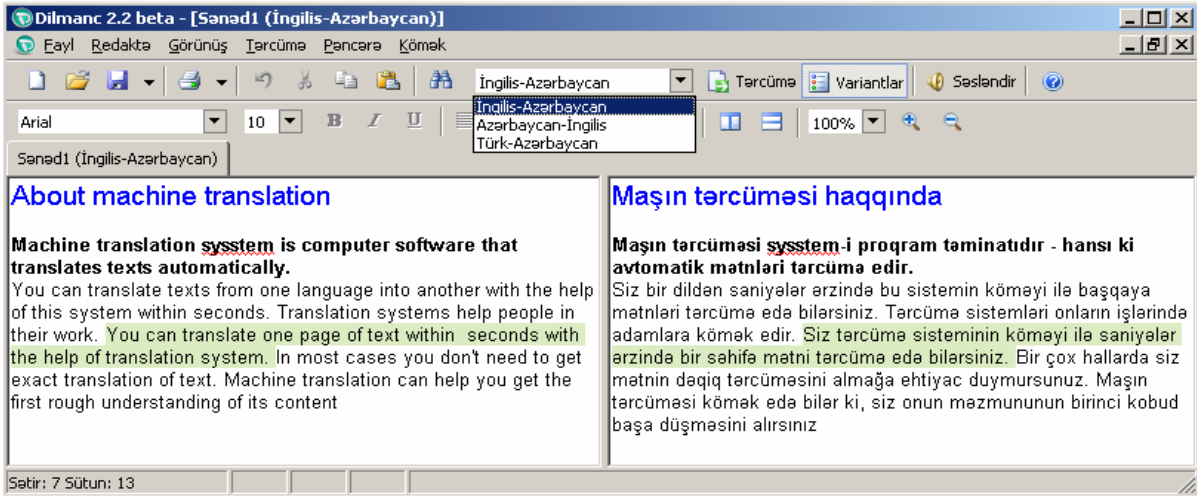

Figure 1: Dilmanc MT system

basis of RBMT (Rule Based MT) and SBMT (Statistic Based MT) approaches. Because Azerbaijani is an agglutinative language it is not possible to create many components of MT system (MT dictionary, active – frequently used suffix chains, disambiguation of suffixes etc.) without involvement of statistics. Besides "usual" statistics, SB component of MT system is being developed at present by creating the parallel bilingual English-Azerbaijani text corpus consisting of ≈10 000 sentences (as a first approximation).

On the results of the test it is possible to say that the system gives good enough - intelligible translations in most cases (Fatullayev et al. 2008, http://www.science.az/cyber/pci2008/1.htm).

Because the volume of the paper does not allow explaining of theoretical aspects of the automation of translation process we want to note some numerical characteristics of Dilmanc MT system:

Azerbaijani-English direction.

1. MT dictionary of Azerbaijani word stems (≈120000 lexical units);
2. Active suffix chains database (≈1000 chains);
3. Formalized rules for the lexical and syntactical disambiguation in Azerbaijani (≈1500 rules);
4. Database of translations of the active suffix chains of Azerbaijani (≈2300 rules);
5. Database of the formal signs of the parts of the sentence in Azerbaijani (≈2000 signs);

English-Azerbaijani direction.

1. English-Azerbaijani MT dictionary (≈115000 lexical units);
2. Database of the formalized rules for the lexical and syntactic disambiguation (≈1400 rules);
3. Formalized rules for the synthesis of Azerbaijani suffix chains (≈300 rules);
4. Database of the rules for the delimitation of the homogeneous parts in English (≈90 rules);
5. Database of the rules for delimitation of clauses in the English sentence (≈40 rules);

Turkish-Azerbaijani direction.

1. Turkish-Azerbaijani MT dictionary (≈20000 lexical units);
2. Database of the equivalency of Turkish and Azerbaijani suffix chains (≈1000 chains).

Besides these means the detailed help system is also developed.

This list is only a small part of all algorithmic and non-algorithmic means developed within the frame of the Dilmanc MT system project.

3 Tools of Dilmanc MT system

Because MT systems give draft translation, it is important to work at this draft text in order to get more accurate translation. Editing the translated text can also become time-consuming when MT system's interface does not provide users with appropriate tools.

Dilmanc MT system has three types of tools for working with the translated text: standard editing tools, tools for improving the translation, tools for the insonation of the text in English (Creation of the software for insonation of the Azerbaijani texts is going on at present time).

Standard editing tools exist in all text editors and they are the following: cut, copy, move, paste, format, open, print, save etc (These tools give possibility to users to edit the text without leaving the translator).

Tools for improving the translation are the following: highlighting the original sentence when user clicks on translated sentence and vice versa, possibility of choosing concrete meaning of the word which has multiple meanings, marking the words which are not included in the MT dictionary, applying original text's format to its translation, creation of user dictionary etc (Fig. 1).

In addition, technologies developed within the Dilmanc project can be used while developing other linguistic technologies (speech and other NLP systems) for Azerbaijani. Note, the researches are carried out for Azerbaijani there is no doubt that developed technologies are also applicable for other Turkic languages.

References

Fatullayev R, Abbasov A, Fatullayev A. 2008. *Set of active suffix chains and its role in development of MT system for Azerbaycani.* In: Proc. of IMCSIT-08, Wisla, Poland, pp.363-368

Fatullayev R, Abbasov A, Fatullayev A. 2008. *Peculiarities of the development of the dictionary for the MT System from Azerbaijani.* In: Proc. of EAMT-08, Hamburg, Germany, pp.35-40

Fatullayev R., Mammadova S., Fatullayev A. 2008. *Statistical analysis of the factors influencing the translation quality of the Dilmanc MT system*. In Proc. of PCI-2008, Baku, vol. 1:96-99

Abbasov A, Fatullayev A. 2007. *The use of syntactic and semantic valences of the verb for formal delimitation of verb word phrases.* In: Proc. of L&TC'07, Poznan, Poland, pp. 468-472

Fatullayev A. 2005. *Modeling the translation process and determination of components of the MT system from Azerbaijani into English on the basis of this model.* Transactions of Academy of Sciences of Azerbaijan, Vol. 25(2): 181-186