Peculiarities of the Development of the Dictionary for the MT System from Azerbaijani

Rauf Fatullayev[1], Ali Abbasov[2], Abulfat Fatullayev[2]

[1] National E-Governance Project, Z.Aliyeva str., AZ1000 Baku, Azerbaijan
[2] National Academy of Sciences, Huseyn Cavid str. 17, AZ1143 Baku, Azerbaijan
fatullayev@gmail.com, ali@dcacs.ab.az, fabo@mail.ru

Abstract. Azerbaijani (The Azerbaijani language) is one of the languages of Turkic group which are morphologically rich languages. While creating the machine translation (MT) system from Azerbaijani, it is not possible to create an MT dictionary consisted of all word-forms of Azerbaijani, because Azerbaijani is an agglutinative language and it is possible to generate practically "endless" number of word-forms by using the various suffix chains (compound suffixes consisted of simple suffixes). For the solution of this problem the approach which considers keeping the word stems and the suffix chains separately is offered. These databases compound together the dictionary of the MT system from Azerbaijani and yield possibility to translate any word-form.

Keywords: Azerbaijani, machine translation system, MT dictionaries, automation of the translation process, Turkic languages

1 Introduction

Azerbaijani is one of the languages of Turkic group and these languages form one of the largest language groups (modern Turkish, Azerbaijani, Kazakh, Uzbek, Turkmen, Tatar, Kirghiz and others). Despite some theoretical research, the most of these languages are still less investigated languages (except modern Turkish [1-2]). Turkic languages are the morphologically rich languages and these languages are characterized by highly productive morphology that may produce a very large number of word forms for a given stem [3-5]. If we take into account only verb stems (≈ 8000) in Azerbaijani [6] and minimal number of active (frequently used) suffix chains which can be joined only the verb stems (≈ 500), we would keep 4000000 verb forms in the dictionary of the MT system. If we take into account word-forms which can be formed from the word stems (≈70000 lexical units, including terms [7]) of the other parts of speech the size of the dictionary will be many times more than this number. So, modeling each word-form as a separate lexical unit leads to a number of problems while developing the formal linguistic technologies (machine translation, speech

This research is carried out within the joint project "Development of the Azerbaijani-English MT system" of The Ministry of ICT of Azerbaijan and UNDP-Azerbaijan

recognition, text to speech etc. systems). Including all word-forms to the dictionary of the MT system can not be considered as a rational way.

In this paper we show the way to solve the problem concerning the development of the dictionary of the MT system from Azerbaijani. For this purpose, we create the dictionary of the word stems of Azerbaijani, the database of the suffix chains and the database of the translation rules of the suffix chains. Using these databases we can construct the translation of any Azerbaijani word-form. These three databases compound the dictionary of the MT system from Azerbaijani (This problem is characteristic for the MT dictionaries from Turkic languages into languages not belonging to this group).

Hereinafter Azerbaijani is taken as a source and English as a target language.

2 Suffix chains of Azerbaijani

Despite the great number of suffix chains we should also take into account that the frequency with which all these suffixes are used is not the same. If this fact is considered while creating an MT system, then the difficulties related to the great number of suffix chains can be avoided. Thus not all suffix chains, but the suffix chains used in texts can be determined and included in the database of suffixes and suffix chains (Such research is already carried out and the database consisting of ≈1000 active suffix chains is formed). Fragment of the active suffix chains database is indicated in the Table 1.

Table 1. Database of suffixes and suffix chains (fragment)

Suffix chain	Other variants of suffix chains	Structure of chains	Code of chains
a	*ə, ya, yə*		*003*
acaq	*acağ, əcək, əcəy, yacaq, yacağ, yəcək, yəcəy*		*026*
ılmiş	*ilmış, ulmuş, ülmüş*	*il-miş*	*057061*
ım	*im, um, üm*		*086*
ımda	*imdə, umda, ümdə*	*ım-da*	*086079*
lar	*lər*		*004*
larım	*lərim*	*lar-ım*	*004086*
mirlər	*mırlar, murlar, mürlər*	*m-ir-lər*	*037035004*
...			

Note: The meaning of the 4th column of the table is explained below.

Most of suffix chains in Azerbaijani have more than one variant of writing, i.e. there are the suffix chains with the same function but different spelling (for example, the suffix -*acaq* has seven forms of spelling else - *acağ, əcək, əcəy, yacaq, yacağ, yəcək, yəcəy*). So, when we speak about any suffix chain we consider the set of suffix chains with the same function but different spelling. On the other hand, in formal analysis process, we usually operate not only with concrete word-forms, but also with the set of word-forms with some common signs. For example, one of the grammatical rules of Azerbaijani runs as follows: for the formation of Future simple of any verb in

Azerbaijani it is necessary to add one of the suffixes *-acaq, -əcək, -yacaq, -yəcək* to the end of the stem of the verb (*al-acaq* – he will buy, *gəl-əcək* – he will come, *oxu-yacaq* – he will study, *bellə-yəcək* – he will spade). Therefore, there is a necessity to designate all verbs with some common symbol. It also concerns to other parts of speech and suffix chains. For these reasons we will use some formalism.

We use the digital codes as the common symbols and assign the same code to all variants of the suffix chains with the same meaning. For example, the code *026* signifies all eight variants of the suffix *-acaq*. In other words the code *026* signifies the set of all variants of the suffix *–acaq*. Therefore we can write the following (*{•}* signifies the set):

026 = {acaq, acağ, əcək, əcəy, yacaq, yacağ, yəcək, yəcəy}.

For other suffix chains we can write the similar equations too. Some of such codes are indicated in the 4th column of the Table 1 and hereinafter we will call these codes as suffix chains too. For example, it is possible to write[2] some rows of this table as:

003 = {a, ə, ya, yə};
004086 = {lar-ım, lər-im};
037035004 = {m-ır-lar, m-ir-lər, m-ur-lar, m-ür-lər}.

3 Translation rules of suffix chains

Suffix chains form a set of word-forms joining to different word stems. For example, suffix *004086* can form the noun word-forms in plural joining to any noun stem (*kitab-lar-ım* – my books, *çiçək-lər-im* – my flowers etc.). For the formalization of this fact we also assigned numerical codes to the parts of speech (Table 2).

Table 2. Codes of parts of speech

Part of speech	Code of part of speech
Verb	*001*
Noun	*002*
Pronoun	*003*
Adjective	*004*
Adverb	*005*
Numeral	*006*
…	

As in the case of suffix chains every code signifies the set, for example, the code 001 signifies the set of all verbs, the code *002* signifies the set of all nouns etc. Using these codes we can code all word-stems of Azerbaijani (Table 3).

[2] Suffix chains can be enumerated from *001* to *823* (because number of active suffix chains is less than *1000*). But for clearness we use all codes of simple suffixes for the designation of the set of suffix chain.

Table 3. Dictionary of the Azerbaijani-English MT system

Stems	Code of part of speech	English translation
ağac	*002*	tree
çiçək	*002*	flower
get	*001*	go
kitab	*002*	book
mən	*003*	I
oyna	*001*	play
yaşıl	*004*	green
...		

After these agreements we can move to the formalization of the translation of suffix chains. As an example, let us consider the word-form *kitablarım* (my books). Because the stem of this word-form *kitab* $\in$ *002* (*kitab* – book is the stem of the word-form *kitablarım* and belongs to the set of nouns) and *larım* $\in$ *004086* (Tables 1 and 3), we can write[3]:

$$kitablarım \in 002004086.$$

Such presentation of the any word-form we will call *code-word*.

Any word-form consisting of the noun stem and one of the variants of the suffix chain *-larım* belongs to the set *002004086.* For example: the word-form *çiçək-lərim* (my flowers) $\in$ *002004086* too (according to the Tables 1 and 3).

Code-word presents the set of word-forms with the same grammatical signs, so we can use code-word to formalize the translation rule of the set of word-forms which are presented by this code-word. For all the word-forms belonging to the set *002004086* the translation rules are the same:

1. Translate the stem of the word-form (code *002* – noun, *kitab* - book);
2. Take the stem in plural (code *004* – suffix of plurality, books);
3. Add the word "my" before this stem (code *086* – possessive suffix of the first person, my books).

So, if we create the translation rules of all active code-words, then the translation rule of any word-form can be defined with its belonging to some code-word.

In the conclusion, the translation of any word-form from Azerbaijani into English is carried out in the following sequence: 1st - a group of appropriate lexical units, 2nd English translation of the stem according to the grammar rules of English. Thus we can form the database of translations of code-words and this one is in fact the data-base of translation rules of suffix chains. Fragment of this database is shown in below (Table 4).

In the 1st column of this table the code-words are indicated, the symbols in the 3rd column are used while constructing the appropriate grammatical category of the word-form in English (they are not suffixes which must be added to the end of the English word). For example, the symbol "s" in the 3rd column means that for the noun code-word (code-word which is beginning *002*) the stem of the English translation of

[3] Of course, instead of the *002004086* we can write "noun+suffix of plurality+possessive suffix of the first person". But digital codes yields possibility to write it more compactly.

the word-form in Azerbaijani must be taken in plural, but for the verb code-word (code-word which is beginning *001*) the third person in singular. In the 2nd column are indicated the group of lexical units which is written before the grammatical category of the word-form constructed in the previous indent.

Table 4. Translation rules of suffix chains

Code-word	Add before the stem	Grammatical category
002		
002004		*s*
002080	*of (the)*	
002004086	*my*	*s*
001037035004	*they don't*	
004004	*the*	
001032		*s*
001035		*ed*
001024032	*must be*	*ing*
...		

Because, our goal is the formalization of the translation rules of the suffix chains, at the end we want to answer to the question: why do we also use the code of the part of speech of the word stem in the translation process of the suffix chains? Because, some suffix chains subject to the part of speech of the word stem can be translated by different rules. For example, suffix *–lar* (suffix of plurality) in the word-form *kitab-lar* (book-s, 002004) is translated into English with the suffix of plurality "-s", but in the word-form *yaşıl-lar* (the green, 004004) without the suffix of plurality.

Now let's consider how this database is used in the translation process. This example describes the process only schematically. In fact, wider information is used in translation process of the word-form.

Example. Let's consider the translation process of the word-form *oynamırlar (oyna-m-ır-lar)*. By using the tables 1 and 3 we can write:

$$oynamırlar \in 001037035004.$$

On the Table 4 we can construct the translation of this word-form in English:

$$oynamırlar \Rightarrow they\ don't\ play.$$

Translation process of the word-form is presented graphically in the Fig. 1.

Conclusion

Development of the NLP systems for the Azerbaijani language [8] on the basis of the theoretical results has been carried out since 2003 [9-10]. The approach explained in this article to the development of the dictionary of the MT system from Azerbaijani yields possibility to keep in the dictionary minimal number of the word-forms. There are about 123000 records in all three databases of the Dilmanc MT

system which compound the dictionary of this system (www.dilmanc.az): In the databases of stems – about 120000 (including word phrases), in the databases of active suffix chains – about 1000 chains, in the databases of the translation rules of the suffix chains – about 2000 rules.

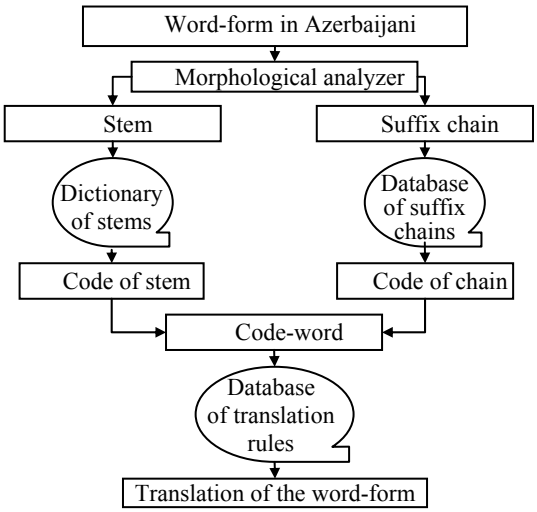

Fig. 1. Translation process of Azerbaijani word-form

The approach explained in this article is also applicable for other Turkic languages and leads to the solution to the problem of the development of the machine dictionary from the Turkic languages.

References

1. Cicekli, I., http://www.cs.bilkent.edu.tr/~ilyas/pubs.html
2. Oflazer, K., http://people.sabanciuniv.edu/oflazer/pubs.html
3. Iskhakova, Kh.F.: Avtomaticheskiy sintez form sushestvitelnogo v tatarskom yazike. Sovetskaya tyurkologiya, 2(8): 20--27 (1968) (in Russian)
4. Bektayev, K.: Statistika kazakhskogo teksta. Gilim, Almaati (1990) (in Russian)
5. Mahmudov, M.: Metnlerin formal tehlili sistemi. Elm, Baku (2002) (in Azerbaijani)
6. Mahmudov, M., Fatullayev, A.: Reverse dictionary of Azerbaijani. Nurlan, Baku (2004) (in Azerbaijani)
7. Akhundov, A. (ed): Spelling dictionary of Azerbaijani. Lider, Baku (2004) (in Azerbaijani)
8. Abdullayev, A., Seyidov, Y., Hasanov, A.: Modern Azerbaijani. Maarif, Baku (1972) (in Azerbaijani)
9. Abbasov, A., Fatullayev, A.: The use of syntactical and semantic valences of the verb for formal delimitation of verb word phrases. In: Proceedings of the 3rd Language & Technology Conference (L&TC'07), pp. 468—472. Poznan, Poland (2007)
10. Fatullayev A., Mehtaliyev A., Ahmedov F., Fatullayev R. Computer translation system from Azerbaijani language into English. Proc. of the 4th international conference Internet-Education-Science, vol. 2, p. 572. Vinnitsia (2004)