STAGES OF THE DEVELOPMENT OF PERSIAN-AZERBAIJANI MT DICTIONARY

Abulfat Fatullayev[1], Niloufar Sadighian[2]

Cybernetics Institute of ANAS, Baku, Azerbaijan
[1]*fabo@box.az*, [2]*n.sadighian@yahoo.com*

Abstract. Bilingual corpora are the important resources in constructing the machine translation systems. In this paper the design process of the monolingual Persian corpus and its use in development of Persian-Azerbaijani MT system are described. Some problems faced experimentally in developing the Persian-Azerbaijani MT system are discussed.

1. Introduction

Persian (also known as Farsi) is one of the Indo-European languages. Unlike many of the Indo-European languages (for example English) it is not written in the Latin alphabet, but uses same script as Arabic, with four additional characters. The Persian language is also one of low investigated languages from the point of view of existence of the NLP systems (MT, ASR, TTS etc.) developed for this language.

Approaches to the development of the modern NLP systems are based on statistical analysis of corpora (monolingual or bilingual), because it gives possibility to detect many regularity of any natural language. Corpora also has great important from the point of view of Machine Translation (MT system). For example, one way to improve the result of automatic translation is to limit the efforts to a specific domain, containing texts of a similar content. The use of corpora and statistics gives possibility to solve the ambiguity problems in the translation process, since many words have much fewer possible translations in a restricted domain [1-4]. MT systems using this approach are often corpus-based, i.e. they have used translation data in the form of parallel corpora to construct their linguistic resources. On the other hand, corpora allow creating the MT (and traditional) dictionaries more relevant to current state of the language. There are several resources in written text that can be used for creating corpora. Newswires and books are the most well-known resources for this task. Nowadays web pages are also widely use as a rich source to construct corpora; because it is possible to collect various texts being a representative of the language by providing the texts in various genres and various authors. The result is providing a reasonable accurate picture of the entire language in which we are interested. Reaching the goal is not an easy task. While working to build a corpus, you might face difficulties; and before processing the corpus, these problems should be removed at a step named "preprocessing". In the experiences made to a Persian-Azerbaijani corpus, we faced a lot of problems due to some special features of Persian. In this paper, we discuss these problems to give a comprehensive perspective of Persian-Azerbaijani corpus to readers.

2. Constructing The Persian corpus

Because, our goal is converting from Persian into Azerbaijani preserving the meaning, we want to answer this question: Why do we also use corpus in the translation process? Firstly it's necessary to identify high occurrence frequencies words from the other in Persian documents secondly there are some polysemous words that should be found the best translation equivalents of them, so parallel corpora statistically provide an alternative solution for overcoming this problem [5-7].

Documents of corpus collection are actually news articles of newspaper (www.hamshahri.net) , scientific sites (www.roshd.ir) , Economic, social, politic, sports and etc articles that is written by many authors from a variety of backgrounds and covers a range of different topics, each with a credible size of data. It also consists of real text in everyday use of

Persians that implies it has the sampling and representativeness property. In addition it used a standard Persian text collection prepared by [8]. The original downloaded documents were HTML files. Then they built a consistent corpus of 110 millions words from those raw HTML files using conversions like: Merging HTML files, removing HTML tags, aligning Persian words, removing numbers, names and punctuations and finally removing pictures. Furthermore, we merged all files to one big file and changed it to readable file by C#.Net. Then words were extracted from file and were counted frequency of each word, they implied to corpus table thus it has 418000 words and their number of frequencies. Because corpus were gathered from varies sites and authors with variety backgrounds, important point in this corpus is that there are all words that usually use in formal written &spoken but there aren’t some words that use in informal spoken such as"تابلو " or "سه" which used instead of the "واضح" (plain) or "خراب" (lost).

In other case second table of corpus has 3000 phrases and idioms that extracted from [9-10] and usually use in formal sentences. Surely the processing of texts out of these sources beside other sources would be a real big challenge.

Total word in the downloaded files after aligning and changing to readable files were 110 millions words. Persian - Azerbaijani Corpus Database held about 415 MB of storage, first table (corpus1) included 418000 words; second table (corpus2) engrossed 3000 phrases and idioms. The Highest frequency Persian word is "و" with 6 million times repetition and after that 2 or 3 letters words are high frequency (shows in table 1). Long words usually have low frequency and most of the valid them (longer than 9 letters) came from other languages such as English or French and or typists made mistake.

Table 1. Table of frequency of Persian word-forms

Word-forms	Azerbaijani	English	Frequency
و	və	and	6,327,214
در	də/da	in/on/at	4,778,642
از	dan/dən	of/from/than/since	4,410,258
به	la/ə	to/at/by/with	3,986,152
را	na/dan	in/of	2,121,514
است	dir/dur	is	2,025,805
…	…	…	…
کم‌آزارترین	ən zərərsiz	the most harmless	9
بسیارخشمگین	çox acıqlı	very angry	9
فدراسیون‌بین‌المللی	beynəlxalq federasiya	international federation	4
نئورئالیسم	yeni realizm	new realism	2
نانوتکنولوژی	nanotexnologiya	nanotechnology	2
گالینابلانکا	Galina Belanka	Galina Belanka	1

The Purpose of this paper is to create a MT system that can translate Persian documents into Azerbaijani documents, so the corpus should be including Azerbaijani Translation of Persian words. Because the words with low occurrence frequency have used lower than high occurrence frequency words in Persian documents, the words with minimum 100 times repetition were chosen for writing Azerbaijani Translation and total number of them are 21000 words. In addition, there are 3000 Persian phrases and idioms into second table of the corpus that should have been write Azerbaijani Translation for them.

One program was written in C#.NET that it can accept up to 3 Azerbaijani words Translation for each Persian word. This program let us add, edit or delete records and it can do sort and search them with various criteria. The corpus was completed by this program, thus it includes 24 thousands Persian words and phrases with maximum 3 Azerbaijani Translation [11] for apiece.

3. Experimental Results and Discussion

While creating Persian-Azerbaijani corpus from written text, we faced several problems based on specific features of the Persian language that will mention in continues. In this section the problems we faced in our experience while constructing a corpus are listed and the possible solutions to fix these problems are discussed.

- There are some abbreviations in Persian documents that are presented with one-letter words. For example, " ع" stands for "علیه السلام" or "ص"instead of the ,"صلی اله"in addition these are Persian letters. In the first situation these words can be written with other names, but in opposite situation they are written as a single word**.**

- In Persian language there are different ambiguities in the lexicon. Because in Persian text short vowels not written but it is pronounced, one of these ambiguities is homographs features. Any morphological analyzers for Persian should be able to recognize and disambiguate such words. For example the homograph "لنگ"could be pronounced any of these along with their POS tagging: /long/ 'Apron' (noun) and /lang/ 'limp' (adj), /leng/ 'foot' (noun). On the other hand, it makes problems in tokenization process to extract the frequency, because these three words would be counted as one, and the result is an unreliable statistics for the frequency of such words. Therefore there is a need to disambiguate the ambiguous words that occur during the translations [12].

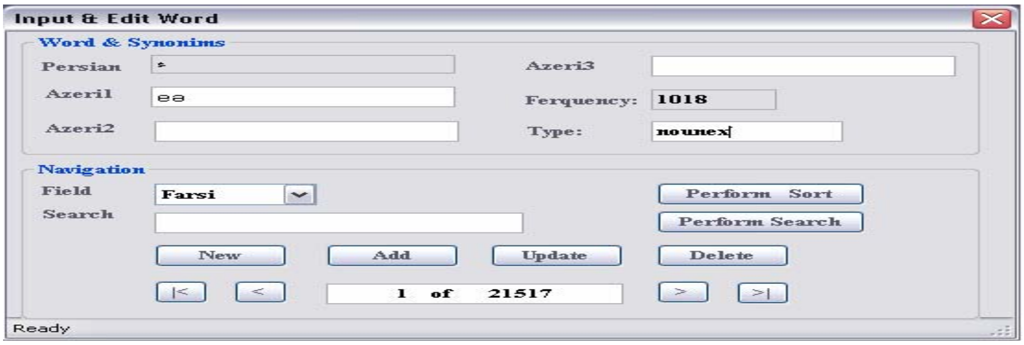
- Another problem in Persian text like any other languages is homonym, since these words pronounced same as each other they are distinguished just in sentences in relation of other words. For example the word "داد" /dad/ is the verb meaning "gave" (verb past tense);"justice" (adj), and it also means "shout" (verb). These words could be disambiguated based on the local context they are used in.

- The optional type of some characters make the processing of a corpus more challenging, because the same words can appear with different forms. One of these characters that borrow from Arabic is "ئ"/-e/. Some typist use "ئ" /-e/ and some "ی" /ye/ instead it. This different typing style causes to have two separated words in the corpus for a single word like "وسائل" /vasa_el/ and "وسایل"/vasayel/ meaning of both of them is 'utensil'.

4. Conclusions

In this paper we have given an overview of design Persian-Azerbaijani MT dictionary. For this reason we created a corpus included Persian word-forms, phrases, idioms and their Azerbaijani translations. Finally we mentioned some of the problems that come from Persian text features and described our solutions to fix them. The volume (number of dictionary entries) of the first version of Persian-Azerbaijani MT dictionary is about 21,000 words and 3000 phrases.

Currently the works on creation of Persian-Azerbaijani MT system are continued.



Input & Edit Word

Word & Synonims

Persian: ء

Azeri1: eə

Azeri2:

Azeri3:

Ferquency: 1018

Type: nounex

Navigation

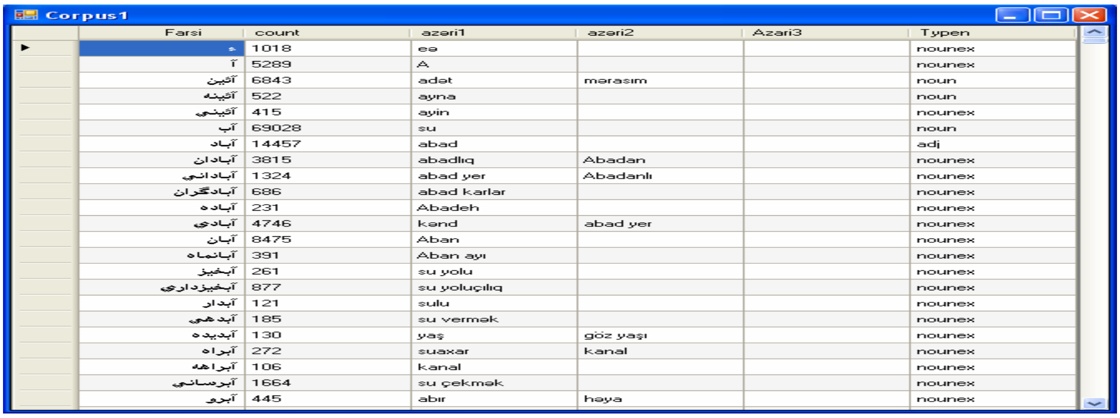
Field: Farsi

Search:

1 of 21517

Ready

Fig.1. Persian-Azerbaijani MT dictionary editor



	Farsi	count	azəri1	azəri2	Azari3	Typen
▶	ء	1018	ee			nounex
	آ	5289	A			nounex
	آئين	6843	adət	mərasim		noun
	آئينه	522	ayna			noun
	آئيني	415	ayin			nounex
	آب	69028	su			noun
	آباد	14457	abad			adj
	آبادان	3815	abadlıq	Abadan		nounex
	آباداني	1324	abad yer	Abadanlı		nounex
	آبادگران	686	abad karlar			nounex
	آباده	231	Abadeh			nounex
	آبادي	4746	kənd	abad yer		nounex
	آبان	8475	Aban			nounex
	آبانماه	391	Aban ayı			nounex
	آبخيز	261	su yolu			nounex
	آبخيزداري	877	su yoluçılıq			nounex
	آبدار	121	sulu			nounex
	آبدهي	185	su vermək			nounex
	آبديده	130	yaş	göz yaşı		nounex
	آبراه	272	suaxar	kanal		nounex
	آبراهه	106	kanal			nounex
	آبرساني	1664	su çekmək			nounex
	آبرو	445	abır	haya		nounex

Fig. 2. Inside of the dictionary

References

[1] Rauf Fatullayev, Ali Abbasov, Abulfat Fatullayev "Peculiarities of the Development of the MT System from Azerbaijani" 12th EAMT conference, 22-23 September 2008, Hamburg, Germany.

[2] Raghavendra Udupa, K.Saravanan, A.Kumaran, Jagadeesh Jagarlamundi "A method for effective and scalable mining of named entity transliterations from large comparable corpora". 12th Conference of the European Chapter of the ACL, Athens, Greece, 30 March – 3 April 2009; pp. 799-807.

[3] Els Lefever, Lieve Macken & Veronique Hoste "Language-independent bilingual terminology extraction from a multilingual parallel corpus". 12th Conference of the European Chapter of the ACL, Athens, Greece, 30 March – 3 April 2009; pp. 496-504.

[4] Masood Ghayoomi, Saeedeh Momtazi, Mahmood Bijankhan "A Study of Corpus Development for Persian", International Journal on Asian Language Processing 20 (1): 17-33, 2004.

[5] Mirko Plitt & François Masselot "A productivity test of statistical machine translation post-editing in a typical localization context". 25-30 January, Dublin, Ireland; Prague Bulletin of Mathematical linguistics, no.93, January 2010; pp. 7-16.

[6] Mohand Beddar, "French to Arabic machine translation: isomorphic syntax, use of terminal sequences", ISMTCL: International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages, and their application to emergencies and safety critical domains, July 1-3, 2009, Centre Tesnière, University of Franche-Comté, Besançon, France (Presses universitaires de Franche-Comté, 2009); pp. 38-42.

[7] Sadaf Abdul-Rauf & Holger Schwenk, "On the use of comparable corpora to improve SMT performance", EACL-2009, Proceedings of the 12th Conference of the European Chapter of the ACL, Athens, Greece, 30 March – 3 April 2009; pp. 16-23.

[8] Abolfazl AleAhmad, Hadi Amiri, Ehsan Darrudi, Masoud Rahgozar, Farhad Oroumchian, "Hamshahri: A standard Persian text collection", Journal of Knowledge-Based Systems, Vol. 22 No.5, p.382-387, Elsevier, July, 2009.

[9] Abdollah Ghanbari, "Persian –English Idioms & expressions Dictionary", Rahnama, 2009, 908 p.

[10] S. Haim, "Persian-English & English-Persian Dictionary", Moaser, 2007, 1317 p.

[11] Möhsün Nağısoylu, "Persian-Azerbaijani Dictionary ", Baku, 2007, 634 p.

[12] Farag Ahmed and Andreas Nürnberger, "Corpora based Approach for Arabic/English Word Translation Disambiguation", Speech and Language Technology, Volume 11, pp. 195-214, 2009.