TEP: Tehran English-Persian Parallel Corpus

Mohammad Taher Pilevar[1], Heshaam Faili[1], and Abdol Hamid Pilevar[2]

[1] Natural Language Processing Laboratory,
University of Tehran, Iran
{t.pilevar,h.faili}@ut.ac.ir
[2] Faculty of Computer Engineering,
Bu Ali Sina University, Hamedan, Iran
pilevar@basu.ac.ir

Abstract. Parallel corpora are one of the key resources in natural language processing. In spite of their importance in many multi-lingual applications, no large-scale English-Persian corpus has been made available so far, given the difficulties in its creation and the intensive labors required. In this paper, the construction process of Tehran English-Persian parallel corpus (TEP) using movie subtitles, together with some of the difficulties we experienced during data extraction and sentence alignment are addressed. To the best of our knowledge, TEP has been the first freely released large-scale (in order of million words) English-Persian parallel corpus.

1 Parallel Corpora

Text corpus is a structured electronic source of data to be analyzed for natural language processing applications. A corpus may contain texts in a single language (monolingual corpus) or in multiple languages (multilingual corpus). Corpora are the main resources in corpus linguistics to study the language as expressed in samples or real world text. Parallel corpora are specially formatted multilingual corpora whose contents are aligned side-by-side in order to be used for comparison purpose.

While there are various resources such as newswires, books and websites that can be used to construct monolingual corpora, parallel corpora need more specific types of multilingual resources which are comparatively more difficult to obtain. As a result, large-scale parallel corpora are rarely available especially for lesser studied languages like Persian.

1.1 Properties of Parallel Corpora

Parallel corpora possess some properties that should be taken into account in their development [1]. The first feature is the structural distance between the text pair which indicates whether the translation is literal or free. Literal and free translations are two basic skills of human translation. A literal translation (also known as word-for-word translation) is a translation that closely follows the form of source language. It is admitted in the machine translation community that the training data of literal type better suits statistical machine translation (SMT) systems at their present level of intelligence [2].

A. Gelbukh (Ed.): CICLing 2011, Part II, LNCS 6609, pp. 68–79, 2011.

The second feature is the amount of noise available in the text pair. Noise is defined as the amount of omissions or the difference in segmentations of the text pair. Another important feature of a parallel corpus is its textual size. The value of a parallel corpus usually grows with its size and with the number of languages for which translations exist. Other features include typological distance, error rate and acceptable amount of manual checking.

1.2 Previous Parallel Corpora for Persian

Persian (locally called Farsi) is an Indo-Iranian branch of the Indo-European languages which uses a modified Arabic script and is spoken in Iran, Afghanistan, Tajikistan, by minorities in some of the countries in the south of the Persian Gulf, and some other countries. In total, it is spoken by approximately 134 million people around the world as first or second language[1]. It is written from right to left with some letters joined as in Arabic. Persian is a highly inflective language in which a great number of different word-forms are created by the attachment of affixes. Persian is a null-subject, or pro-drop language, so personal pronouns (e.g. I, he, she) are optional.

Until the release of TEP, there were quite a few parallel corpora for Persian language which were either small in size or unavailable for research purpose. Lack of such a resource hindered research in multilingual NLP applications such as statistical machine translation for Persian language.

Shiraz corpus is a bilingual parallel tagged corpus consisting of 3000 Persian sentences with the corresponding English sentences. The corpus is collected from Hamshahri newspaper online archive and all its sentences are manually translated at CRL3 of New Mexico State University [3].

In [4], in order to train a Persian-English speech to speech translation device, the authors have collected a corpus of medical-domain parallel cross-lingual transcripts which is composed of about 300K words.

The authors in [5] present a method to create Persian-English sentence-aligned corpus by mining Wikipedia. They used Wikipedia as a comparable corpus and extracted aligned sentences from it to generate a bilingual parallel corpus. They ran the method on 1600 page pairs which yielded about 12530 sentence pairs. The resulting corpus, however, has not yet been released.

In [6], Miangah reports an attempt to constitute an English-Persian parallel corpus composed of digital texts and web documents containing little or no noise. The corpus consists of total 3.5M English and Persian words aligned at sentence level (about 100K sentences, distributed over 50,021 entries). Although the corpus seems to have been offered in ELRA's website[2], we could not obtain a copy of it for academic purpose. Upon our inquiry, the developers expressed their unwillingness to release the corpus for it being still under development.

1.3 Using Movie Subtitles to Construct Parallel Corpora

The first resource that usually comes under consideration for construction of parallel corpora is literary translations. They are however less common for machine

[1] Languages of the World, 2005.
[2] `http://catalog.elra.info`

translation purpose, because they do not usually adopt literal translations and therefore involve many content omissions. This non-literal type of translation does not suit the word alignment process which is an essential step in the training of statistical machine translation systems. Translated books are not only unsuitable for the purpose, but also protected by copyright. Literal translations such as Hansards are commonly used in MT community as a resource to generate parallel corpora. For European languages, the Europarl corpus has become quite a standard one. Unfortunately, there exists no similar resource for Persian language.

To acquire a fairly large parallel corpus that could provide the necessary training data for experiments on statistical machine translation, we chose to mine movie subtitles; a resource which until recently has not been utilized by NLP tasks. There are various advantages in using movie subtitles [7], such as:

- They grow daily in amount: due to high demand, the online databases of movie subtitles are one of the fastest growing multilingual resources.
- They are publicly available and can be downloaded freely from a variety of subtitle websites.
- The subtitle files contain timing information which can be exploited to significantly improve the quality of the alignment. Fig. 1 shows a small part of a movie subtitle file.
- Translated subtitles are very similar to those in the original language – contrary to many other textual resources; the translator must adhere to the transcript and cannot skip, rewrite, or reorder paragraphs.

Fig. 1. A manually aligned part of a movie subtitle pair

There are however disadvantages to using movie subtitles as a bilingual resource:

- Movie subtitles typically contain transcriptions of spontaneous speech and daily conversations which are informal in nature, and therefore the output of a machine translation system trained on them will be biased towards spoken language.

- After investigating the translated sentences in a statistical machine translation trained on an English-Persian corpus of movie subtitles [8], we observed that the average sentence length ratio of Persian to English is about 0.7 (which is not the case in human translation). This means that this resource is not well-suited for machine translation purpose.
- Punctuations are not usually included in movie subtitles, and therefore sentence limits are not available. This is especially problematic for a language like Persian whose sentences do not begin with a capital letter or a similar distinctive feature. For movie subtitles, the alignments are usually made between individual lines in subtitle files according to the timing information. However these individual lines are sometimes neither complete sentences nor complete phrases. This in turn leads to several problems. In 3.4.1, we will discuss some more problems faced while constructing parallel corpora from movie subtitles.
- In Persian, words are spoken in many ways, and therefore written in many different forms in an informal text like movie subtitle. Unifying these forms to avoid the scarcity is to be done manually and needs great effort.

Some of these problems can be tackled by applying rule-based correction methods. Building aligned bilingual corpora from movie subtitles were first presented in [9]. They proposed a semi-automatic method which needs human operator to synchronize some of the subtitles. Tiedemann created a multilingual parallel corpus of movie subtitles using roughly 23,000 pairs of aligned subtitles covering about 2,700 movies in 29 languages [10]. He proposed an alignment approach based on time overlaps. The authors in [7] proposed a methodology based on the Gale and Church's sentence alignment algorithm, which benefits from timing information in order to obtain more accurate results.

2 Persian Informal/Spoken Language

In most languages, people talk differently from the way they write. The language in its spoken (oral) form is usually much more dynamic and immediate than its written form. The written form of a language usually involves a higher level of formality, whereas the spoken form is characterized by many contractions and abbreviations. In formal written texts, longer and more difficult sentences tend to be used, because people can re-read the difficult parts if they lose track. The spoken form is shorter also due to semantic augmentation by visual cues that are not available in written text.

The size of the vocabulary in use is one of the most noticeable differences between oral and written forms of discourse. Written language uses synonyms instead of repeating the same word over and over again. This is, however, not the case in oral language which usually makes use of a more limited vocabulary. The level of difficulty in pronunciation may also affect the words chosen. Oral languages tend to use words of fewer syllables.

In addition to the aforementioned general differences between spoken and written forms of a language, Persian language introduces a variety of differences which further expand this gap. In addition to many informal words not appropriate to be used in formal language, there are remarkable variations in pronunciation of words.

As a case in point, the word “nan” (“a” is pronounced as the only vowel in the word “got” in English), which means bread, is changed into “noon” (“oo” as in “cool”) in spoken language. This alteration between “aa” and “oo” is quite common but has no rule; so in many words speaker is not allowed to interchange “aa” and “oo” in colloquial language. Another common case is changing the last part of verbs. For example, the verb “mi:ravad” (“i:” as in “see” & “a” as in “cat”), which means she/he is going, changes into “mi:reh” (“i:” as in “see” & “eh” as in “pen”).

A subtitle file reflects exact conversions of a movie in written form. A Persian subtitle file, therefore, involves all of described features of the spoken form of Persian language.

3 Development of the Corpus

3.1 Resources

Around 21000 subtitle files were obtained from Open-subtitles[3], a free online collection of movie subtitles in many languages. It included subtitles of multiple versions of the same movie or even multiple copies of the same version created by different subtitle makers. For each movie, a subtitle pair was extracted by examining the file size and timing information of available subtitle files. These information were used to confirm that the subtitle file pair belonged to the same version of a movie. Duplicates were then removed to make the resource unique and avoid redundancy. It resulted in about 1200 subtitle pairs. Each pair comprised of two textual files (in srt format), containing subtitles of the same version of a movie in both Persian and English languages.

3.2 Preprocessing

The movie subtitles database is entirely made up of user uploads and due to lack of a standard checking procedure, they need to be overviewed first. This overview includes checking if movies are tagged with the correct language, or if they are encoded in the same character encoding. We will talk more about this in 3.3.

Out of available subtitle formats, we selected those formatted using two most popular formats: SubRip files (usually with extension ‘.srt’) and microDVD subtitle files (usually with extension ‘.sub’). We then converted files with these formats to a standard XML format.

3.3 Subtitle Alignment

Subtitle alignment is essentially similar to normal sentence alignment. Movie subtitles, however have an advantage as most of alignments are 1:1 and that they carry additional information that can help alignment.

We used the method proposed in [7] which is based on the algorithm proposed by Gale and Church with a small modification in order to take full advantage of timing information in movie subtitles. This method is a dynamic programming algorithm that

[3] www.opensubtitles.org

tries to find a minimal cost alignment satisfying some constraints. According to [11], the recursive definition of alignment cost is calculated by the following recurrence:

$$C(i,j) = min \begin{cases} \text{C(i,j-1)} + & d(0,t_j;0,0) \\ \text{C(i-1,j)} + & d(s_i,0;0,0) \\ \text{C(i-1,j-1)} + & d(s_i,t_j;0,0) \\ \text{C(i-1,j-2)} + & d(s_i,t_j;0,t_{j-1}) \\ \text{C(i-1,j-1)} + & d(s_i,t_j;s_{i-1},0) \\ \text{C(i-2,j-2)} + & d(s_i,t_j;s_{i-1},t_{j-1}) \end{cases} \quad (1)$$

where *C(i,j)* is the alignment cost of a sentence in one language (s_i, *i=1…I*) with its translation in another language (t_j, *j=1…J*). *d(e,f)* is the cost of aligning *e* with *f*. Gale and Church defined *d(e,f)* by means of relative normalized length of sentence in characters, namely $l(e)/l(S_e)$ and $l(f)/l(S_f)$ where $l(S_e)$ and $l(S_f)$ are the total lengths of the subtitle files of the first and second language, respectively. The authors in [7] defined a new cost function that also used the timing information. The specific cost function for subtitle alignment is as follows:

$$d(e,f) = \lambda \left(\frac{dur(e)}{dur(S_e)} - \frac{dur(f)}{dur(S_f)} \right)^2 + (1-\lambda) \left(\frac{l(e)}{l(S_e)} - \frac{l(f)}{l(S_f)} \right)^2 \quad (2)$$

where the duration *dur(s)* of subtitle *s* is defined as:

$$dur(s) = end(s) - begin(s) \quad (3)$$

And λ is a language-dependent parameter whose value can be determined using grid-search and represents the relative importance of the timing information. We used the above algorithm for aligning subtitles using which we were able to produce highly accurate alignments.

3.4 Problems in Corpus Building

3.4.1 Problems with Subtitles

As mentioned earlier in 1.3, there are some disadvantages to making parallel corpora from movie subtitles. Apart from that, we experienced various impediments in extracting the parallel content from subtitle files. Some of these issues are listed below:

1. Noise: most of the subtitle files begin or end with advertising messages, comments about the subtitle creation/translation team or similar content which do not usually match between the file pair. An example is shown in Fig. 2. These non-matching contents of subtitle file pair introduce difficulty while aligning their sentences. There is no straightforward method to tackle this noise and hence, a manual process is required to chop off these contents from subtitle files. The user needs to spend considerable amount of time to remove this kind of noise in a text editing software.

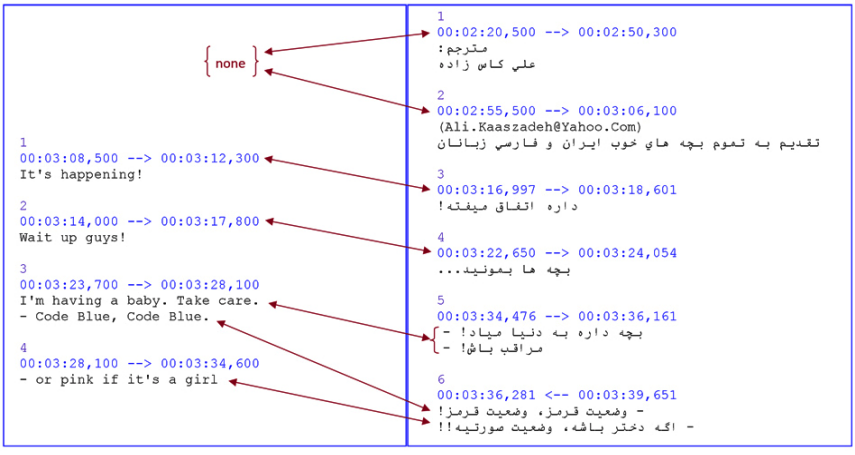

Fig. 2. An example for the available noise in subtitle pairs (advertisement at the beginning of Persian subtitle file)

2. The timing in subtitles is usually specified by frame numbers and not by real time, and therefore the absolute time values cannot be utilized for alignment and converting it to real time values is not always possible. Hence we use normalized subtitle duration instead [7]. This results in a significant reduction in the alignment error rate.
3. Another important problem with the subtitle files as parallel corpus construction resource is that their sentences usually do not end with full stops. As a result, the outputs of alignment phase are chunk pairs rather than sentence pairs. This is not obviously very desirable since it reduces the quality of further processing such as syntactic parsing. This is especially problematic for languages like Persian in which, the sentences do not begin with capital letters or a similar distinct notation and therefore no sentence splitting method is available.
4. Sometimes the frame rates of subtitle pairs do not match. This can however be easily sorted out using available subtitle editing software, but the process of identifying such mismatches is itself time taking.
5. Some of the subtitles are specifically intended for people who are deaf and hard-of-hearing. These are a transcription rather than a translation, and usually contain descriptions of important non-dialog audio such as “[sighs]” or “[door creaks]”. These content sometimes do not match between subtitle pairs and introduce some difficulties while alignment. They are however easy to detect as they usually come within square brackets. This enables a simple regular expression to remove them completely.
6. Subtitle files do not have a standard for encoding. In order to overcome this, we converted all files to UTF-8 encoding.
7. While processing, we figured out that there are many subtitles with incorrect language tags. We used a simple Trigram Technique to detect the language of subtitle files. Therefore, each subtitle file was analyzed during preprocessing to

check whether the specified language tag correctly describes the language of that file. We also ignored those subtitles that contained multiple translations within the same file.

8. Another major drawback with the use of movie subtitles in corpus construction is that the resulting corpus cannot be easily annotated in an automatic manner. We made an effort to generate parse-trees of the sentences in TEP. However we soon realized that the spoken nature of the sentences in the corpus does not allow a reliable parsing. This was also problematic in the case of part of speech tagging. Especially in the case of Persian in which many words are different in spoken and written forms, no simple remedy exists to efficiently generate PoS tags using a PoS tagger trained on the available formal training texts.

3.4.2 Problem with Persian

In this section we report on some problems we ran into while developing English-Persian parallel corpus, most of which originated from specific features of Persian language. We will also discuss possible solutions to tackle some of these problems.

Persian uses code characters that are very similar to that of Arabic with some minor differences. The first difference is that the Persian alphabet adds 4 letters to that of Arabic. The other one is that the Persian employs some Arabic or ASCII characters beside the range of Unicode characters dedicated to it. Hence, the letters ک (kaf) and ي (ye) can be expressed by either the Persian Unicode encoding (U+06A9 and U+064A) or by the Arabic Unicode (U+0643 and U+06CC or U+0649) [12] and [13]. Therefore, to standardize the text, we replaced all Arabic characters with their equivalent Persian characters.

Another problematic issue while processing Persian texts is the internal word boundary in multi-token words that should be presented by a pseudo-space which is a zero-width non-joiner space. Amateur Persian typists tend to type a white-space instead of the pseudo-space in multi-token words. In such a case, a single multi-token word is broken up into multiple words which in turn introduce several problems while processing. For example words such as "می‌شود" and "پایان‌نامه" are sometimes mistakenly typed as "می شود" and "پایان نامه" which are both broken into two independently meaningful words when separated by a space. Obviously, such an issue affects statistical analysis of the text. In Persian, there exist many multi-token words, for which the insertion of pseudo-space is optional. For instance morphemes like the plural morpheme (ها), comparative suffix (تر، ترین) can be either joined or detached from words. This can result in distribution of the frequency of such words between different typing styles. However in a standard Persian corpus, these affixes are very limited in number and do not usually include ambiguities, and therefore a major part of such problems can be overcome.

As mentioned earlier in 2, there usually exist some differences in pronunciation of words between spoken (informal) and written (formal) Persian. As a case in point, the word آتش (pronounced as /ʌtæsh/) which means fire is changed into آتیش (pronounced as /ʌtɪsh/) in informal Persian. This alteration between "æ" and "ɪ" is quite common but has no rule. There are many more cases where a difference between spoken and

formal Persian exists. This phenomenon is not observed in English, except for a few words like “them” which is sometimes written as “em” in spoken language. Table 1 shows examples of such words along with their frequencies in the TEP corpus.

This feature of Persian language has a negative effect on the quality of applications such as word frequency analysis or statistical machine translation when trained on a corpus of spoken language. Our effort in finding a set of rules to efficiently switch Persian words between their spoken and written forms did not result in any concrete way to merge these multiple styles into a unique form. Unlike the case of morphemes that can be automatically resolved, this multi-style issue of Persian cannot be overcome in a straightforward manner. Hence, we tried to manually transfer as many multi-style forms to a unique form as possible.

Table 1. Examples for Persian words having different written styles in formal and informal language (along with their frequencies in TEP corpus)

Spoken form	Freq.	Formal form	Freq.
خوندم (khundam)	194	خواندم (khāndam)	23
آتیش (ātish)	140	آتش (ātash)	972
نمیتونم (nemitunam)	2381	نمي‌توانم (nemitavānam)	56
بهش (behesh)	5674	به او (be oo)	2384

4 Statistics of TEP

Table 2 summarizes the statistics of the first release of TEP. Fig. 3 and 4 show the sentence length distributions of English sentences in characters and words respectively, whereas Fig. 5 and 6 illustrate that of Persian sentences. As observed in Table 2, the number of unique words in Persian side of the corpus is about 1.6 times more than that of English. This is due to the rich inflectional morphology of Persian language. We can also conclude that Persian sentences are on average constructed using fewer characters in comparison to their equivalent English sentences.

Table 2. Statistics of TEP

	English side	Persian side
Corpus size (in words) excluding punctuations	3,787,486	3,709,406
Corpus size (in characters excluding space)	15,186,012	13,959,741
Average sentence length (in words)	6.829	6.688
No. of unique words	75,474	120,405
Corpus size (in lines)	554,621	

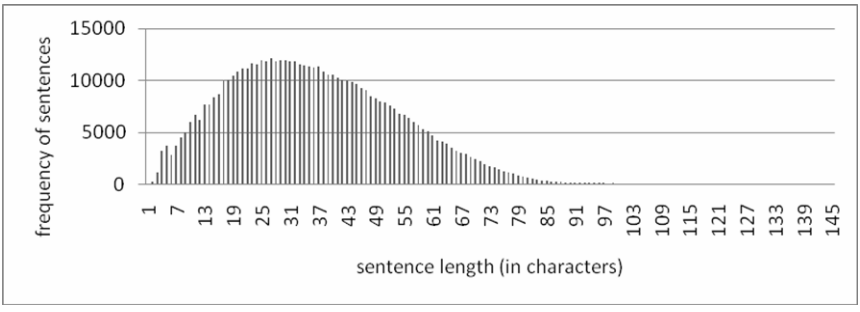

Fig. 3. Distribution of English sentences according to their lengths (in characters)

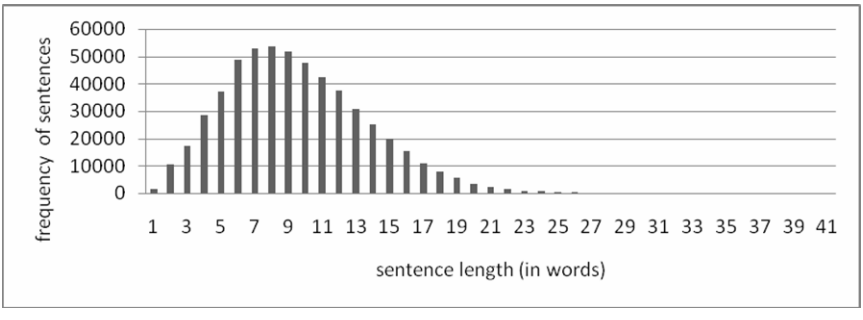

Fig. 4. Distribution of English sentences according to their lengths (in words)

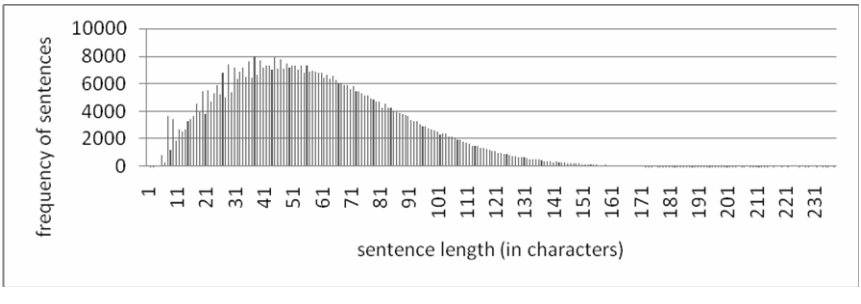

Fig. 5. Distribution of Persian sentences according to their lengths (in characters)

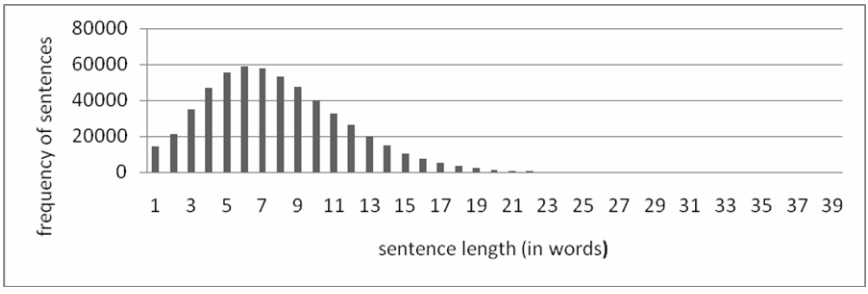

Fig. 6. Distribution of Persian sentences according to their lengths (in words)

5 Release of the Corpus

TEP is released freely under the GPL[4] in Feb. 2010. For more details, please check the website[5] of Natural Language Processing laboratory of University of Tehran. The second release of the corpus is expected to be on 2011. Tehran Monolingual Corpus (TMC) which is the largest available monolingual corpus for Persian language is also available for download at website. TMC is suitable for language modeling purpose.

6 Conclusion

In this paper we described the development of TEP corpus and also mentioned some of the problems faced in parallel corpus construction from movie subtitles together with possible solutions to them. TEP can be advantageous to researchers in several NLP areas such as statistical machine translation, cross-lingual information retrieval, and bilingual lexicography. We hope that our work would bring about more efforts to develop large-scale parallel corpora for Persian language.

References

1. Rosen, A.: Building a parallel corpus for too many language. JRC workshop on exploiting parallel corpora in up to 20 languages (2005)
2. Han, X., Li, H., Zhao, T.: Train the machine with what it can learn-corpus selection for smt. In: 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora, BUCC 2009, pp. 27–33. Association for Computational Linguistics, Morristown (2009)
3. Amtrup, J.W., Mansouri Rad, H., Megerdoomian, K., Zajac, R.: Persian-english machine translation: An overview of the shiraz project. NMSU, CRL, Memoranda in Computer and Cognitive Science (MCCS-00-319) (2000)
4. Georgiou, P.G., Sethy, A., Shin, J., Narayanan, S.: An english-persian automatic speech translator: Recent developments in domain portability and user modeling. In: Proceedings of ISYC 2006, Ayia (2006)

[4] GNU General Public Licensing (`www.gnu.org/licenses/gpl.html`)
[5] `http://ece.ut.ac.ir/nlp/`

5. Mohammadi, M., GhasemAghaee, N.: Building bilingual parallel corpora based on wikipedia, pp. 264–268. IEEE Computer Society, Los Alamitos (2010)
6. Mosavi Miangah, T.: Constructing a large-scale english-persian parallel corpus. META 54(1), 181–188 (2009)
7. Itamar, E., Itai, A.: Using movie subtitles for creating a large-scale bilingual corpora. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008). European Language Resources Association (ELRA), Marrakech (2008)
8. Pilevar, M.T., Feili, H.: Persiansmt: A first attempt to english-persian statistical machine translation. In: Proceedings of 10th International Conference on statistical analysis of textual data, JADT 2009, pp. 1101–1112 (2009)
9. Mangeot, M., Giguet, E.: Multilingual aligned corpora from movie subtitles. Technical report, Condillac-LISTIC (2005)
10. Tiedemann, J.: Improved sentence alignment for movie subtitles. In: Proceedings of Int. Conf. on Recent Advances in Natural Language Processing (RANLP 2007), pp. 582–588 (2007)
11. Gale, W.A., Church, K.W.: A program for aligning sentences in bilingual corpora. In: Proceedings of the 29th Annual Meeting on Association for Computational Linguistics, ACL 1991, pp. 177–184. Association for Computational Linguistics, Morristown (1991)
12. Megerdoomian, K.: Persian computational morphology: A unification-based approach. Cognitive Science (2000)
13. Ghayoomi, M., Momtazi, S., Bijankhan, M.: A study of corpus development for persian. International Journal on Asian Language Processing 20(1), 17–33 (2010)