Research on the Automatic Extraction of Persian Simple Sentences

Wei Li
Luoyang University of Foreign Languages 471003 Luoyang, Henan, China
lw_waiyuan@163.com

Keywords: Persian; Corpus; Clause; Automatic extraction

Abstract. In order to meet the needs of natural language shallow parsing for Persian, this paper analyzes the boundary rules of Persian sentences and puts forward an effective algorithm and program of extracting Persian simple sentences by repeated experiments on the basis of Uppsala Persian Corpus. This algorithm, which can reduce the difficulty of parsing, is of vital importance to information retrieval, machine translation, phrase recognition, the construction of network answering system, etc. Besides, the obtained simple sentences can also be useful resources for Persian linguistic studies.

Introduction
One of the natural language shallow parsing tasks is resolving compound sentences into simple ones, which can lower the difficulty of parsing significantly through simplifying sentence structures. Nowadays, some achievements have been made in the aspect of Persian part-of-speech tagging and there are not only part-of-speech tagging systems of high accuracy, but well-known Persian corpora like Bijankhan Corpus [1] and Uppsala Persian Corpus (hereinafter “UPC”) [2] have been constructed as well. According to the corpora, there are a lot of complex sentences in Persian texts. If the complex sentences can be divided into simple sentences, it is of great significance to both improving processing efficiency and deepening the research on Persian corpora. Therefore, this paper explores characteristics of Persian complex sentences and designs a system of extracting Persian simple sentences on the basis of UPC.

Related Work
Because of the significance of recognizing simple sentences automatically, a lot of researchers at home and abroad have undertaken studies on this aspect. For example, some researchers have discussed the end forms of the Uyghur sentences and realized automatic identification of Uyghur simple sentences by means of the corpus [3]. Some other researchers consider that the clause is the basic unit of Chinese discourses and have implemented a comma-based system of identifying clauses via tagged corpus [4]. Another two scholars have undertaken some studies on topic detection of punctuation sentences by means of the generalized topic theory in order to improve the accuracy of information extraction and machine translation [5]. Still another three think that processing complex sentences correctly in English-Chinese machine translation depends on whether the clauses can be identified and have put forward the detection method of English clauses based on maximum entropy principle [6]. Also, another detection method of English clauses has been discussed in [7] and improvement proposals of adding linguistic rules to the model and modifying smoothing algorithm have been made in the conclusion. [8] has discussed a memory-based method of extracting English clauses, which needs the part-of-speech tags of the words and a base chunk structure of the sentence.

To sum up, simple sentence identification is of great importance in natural language processing and the study of linguistics.

Uppsala Persian Corpus
UPC is the improved version of Bijankhan Corpus. Bijankhan Corpus was released in 2004, which is the first large-scale manually annotated corpus. Its contents are from online novels and newspapers, etc
and have the features of various themes and rich topics. All the words in Bijankhan Corpus are tagged, which contains about 2.6 million tokens. However, as it was constructed not for natural language processing, many features make it not suitable for automatic processing. For example, there exist the problems of lacking sentence segmentation and uneven normalization. Besides, Persian and Arabic letters are mixed together and texts have a mixture of different types of character sets. Therefore, Mojgan Seraji improved Bijankhan Corpus and named the new version UPC. There are two improvements in UPC. For one thing, the tagset of Bijankhan Corpus has been modified via removing, adding and merging categories. For another, mistakes and inconsistencies in the annotation have been corrected. UPC is available at [9]. A screenshot of UPC and its tagset are shown in Fig. 1 and Fig. 2.

![Figure 1. A screenshot of UPC](image1.png)

![Figure 2. The tagset of UPC](image2.png)

### An Algorithm of Extracting Persian Simple Sentences Based on Grammar Rules

According to the Persian grammar, sentences with two or more predicate verbs are called complex sentences, which occupy an important position in Persian expressions. According to the different relationship between clauses, complex sentences can be divided into compound complex sentences and main-subordinate complex sentences. This paper develops the following steps of extracting simple sentences based on the idea of filtering:

First, punctuation marks are applied to make a rough split of complex sentences. In UPC, all the punctuation marks are annotated with “<DELM>”. It can be found by traversing the corpus that clauses are joined mainly by comma, colon and semicolon, which appear as “،” “:” and “;” in Persian.
Through traversing the corpus, I extract every clause joined by comma together with a verb, colon and semicolon and save them as a single line respectively.

Second, lines with no verb are deleted and lines with only one verb are saved as the target simple sentences. The rest are further processed.

Persian modal verbs like “(should)”, “(may)” have gradually been turned into adverbs, which are represented in forms of verbs’ not changing with the subject [10]. The usual forms of “(should)” and “(may)” are “(should)”, “(may)” are “(should)”, “(may)”. Furthermore, when “(should)” and “(may)” are used to make an impersonal sentence, their forms also don’t vary with the subject and they have only three forms “(should)”, “(may)” and “(can)”. In the corpus, all the words mentioned above are annotated as “<V_AUX>”, which means auxiliary verb. Actually, they function as adverbs. Therefore, verbs with “<V_AUX>” are not regarded as a verb. Moreover, two verbs are often joined together in Persian and the form in the corpus is “word<V_AUX>word<V_AUX>”. As they function as predicate verbs together, two verbs which are next to each other are regarded as one verb.

As for “<CON>(and)”, “<CON>(or)” and other conjunctions, if they join two or more verbs, all of them are regarded as one verb.

Third, in this step, a decision should be made which should be processed first, compound complex sentences or main-subordinate complex sentences. The following three examples are used to make an illustration.

1. … و پرویز هم پسری کوشا ست که در روز صبح به میدان ورزش می‌رود و همکلاسهای دیگر بايد از آنها تحسیل می‌کند....

2. (2)چگونه می‌توان به کودکی که نه می‌شنوید و نه حرف می‌زند خواندن و نوشتن آمخت؟

3. (3)همه کشورها اعم از اینکه بزرگ و کوچک باشد، باید متساوی الحقوق باشند.

All countries, big or small, should be equal.

In Example (1), there is a subordinate clause in a compound complex sentence. “و” is a coordinate conjunction, which is the equal of “and” in English. “که” is the relative pronoun of the attributive clause, which is the equal of “that”. If the coordinate conjunction “و” is processed first and I extract compositions until “که”, the clause led by “که” will be separated and the pattern “noun+verb+clause” cannot be found. In Example (2), there is a coordinate clause in a main-subordinate complex sentence. If the clause led by “که” is extracted first, the verb pattern won’t be broken in spite of separating the coordinate clauses from each other. Also, the usage of the conjunction “و” is very common in Example (3). If “و” is processed first, the original structure of the sentence will be broken. Therefore, main-subordinate complex sentences in the corpus should be processed first.

In Persian, main-subordinate complex sentences fall roughly into three categories: noun clauses, adjectival clauses and adverbial clauses. Noun clauses contain subject clauses, object clauses, appositive clauses and predicative clauses. Subject clauses can be divided into preposition subject clauses and post subject clauses according to different positions of the clause, as are shown in Example (4) and Example (5). As Persian belongs to the type of SOV, the verb is preceded by the object. However, there are two situations about the position of object clauses, which can be located either before the verb or after the verb, as are shown in Example (6) and Example (7). Predicative clauses usually come after link verbs, as is shown in Example (8). Adjectival clauses refer to attributive clauses, which are similar to appositive clauses. The difference is that the relative pronoun of attributive clauses is used as a sentence element and the relative pronoun of appositive clauses not. The position of adverbial clauses is very flexible, which can be placed in the initial, medial and end position.

1. (4)هر چه می‌درخشد طلا نیست.

2. (5)معلوم نیست که ساعت چند بر می‌گردد.
We should discuss together what we learn.

Do you know when this game will start?

My first study plan this evening is remembering all the words.

In Example (4), "هر چه" (that glitters) is the subject clause, which is in the initial position and belongs to preposition subject clauses. The relative pronouns are usually "هر چه", "از چه", "هر چه". Example (6) belongs to the situation that the object clause comes before the verb and its relative pronouns share the same ones above. Example (5) and Example (7) stand for post subject clauses and post object clauses, which share the same complementizer "که" with appositive clauses, predicative clauses and attributive clauses. "where", "when", etc in English can be used as relative adverbs and "چه موقع" in Example (7) is the equal of "when" in English, which is not the relative adverb and its position is usually medial. In order to keep as many original sentence elements as possible, take the clause led by "که" for example, sentence elements from "که" to the first verb behind "که" are extracted as a simple sentence and if there is still a verb in the rest of the original sentence, "<C>" is added in the missing position, which represents that a clause can be added here. This method can be named Boundary Word Interception. If the relative pronoun "که" is omitted, the sentence should be disconnected from behind the verb. This method can be named Verbal Disconnection. Here, words like "که" are named the boundary words, which need to contain the part-of-speech tagging information, because sometimes a word has different parts of speech. For instance, "<CON>" is a complementizer, while "<P>" is a preposition. "<CON>" should be used to extract simple sentences instead of "<P>". Moreover, there are two tagging forms about the multi-word boundary word in the corpus. For example, "قبل<CON>" and "پیش<ADV_TIME>از<P>" both appear in the corpus. In order to identify all the boundary words, a corpus-based statistical recognition method is first applied and the result of recognition is shown in Fig. 3 and the three parameters behind the phrases are phrase frequency, phrase probability and the minimum value of number of different words in either the left or the right neighboring place of the possible phrase.

Figure 3. Corpus-based statistical recognition of phrases

The statistically based approach above overcomes the subjective shortcomings and all the boundary words in the corpus are listed after manual verification. Because there are many phrases in the boundary words, whole word matching can’t be applied in the process of identification. However, some boundary words are part of other boundary words. For example, "<CON>" can serve as a boundary word alone, but it is also a part of many other boundary words. Therefore, in order to find all the boundary words, the boundary words need to be sorted first and the sort result is shown in Fig. 4.
After finishing processing the sentences containing boundary words, the rest containing two or more verbs can be divided into two groups: main-subordinate complex sentences with complementizers omitted and compound complex sentences. As the relationships between the clauses of compound complex sentences are equal and “<C>” is not needed, the method of Verbal Disconnection should be used to process compound complex sentences. As for compound complex sentences, the coordinate conjunctions should be identified first via the statistically based approach and manually verification. The result is as follows.

1. <ADV>...<CON>...<CON>...<CON>...<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>,<CON>
Conclusion

This paper aims at keeping as many original sentence elements as possible for the extracted simple sentences. The pattern “verb+clause” is very common in Persian, so this paper adds “<C>” in an attempt to mark the clause position. Although the method Verbal Disconnection can process most Persian complex sentences, yet it can’t distinguish main-subordinate complex sentences and compound complex sentences. However, as the clauses of compound complex sentences have an equal position, “<C>” is not needed. According to the method Verbal Disconnection, as long as there is a coordinate conjunction in a sentence, the sentence should be disconnected from behind the verb without “<C>” and thus the position of the clause can’t be determined. Therefore, Verbal Disconnection is not suitable for multi-complex sentences, which contain both main-subordinate complex sentences and compound complex sentences. In the actual text, multi-complex sentences are very common. Therefore, the method of Boundary Word Interception is also proposed for processing the main-subordinate complex sentences.

The sequence among the steps of the algorithm of extracting Persian simple sentences above is determined through repeated experiments and on the basis of Persian grammar rules. Therefore, almost all the complex sentences in UPC have been processed with one exception that the complementizer of main-subordinate complex sentences is omitted and coordinate conjunctions exist in very few multi-complex sentences. This leads to the loss of “<C>” after the verb. As for UPC, this is a rare case and the correction is done manually, but if the algorithm in this paper is applied to process other Persian corpora and the case above is not rare, the approach of counting occurrence frequency can be used to correct the rare pattern of a verb.

To sum up, good experimental results have proved the effectiveness of the algorithm proposed in this paper. As all the simple sentences are obtained from the authentic corpus, they can also be important resources for Persian linguistic research.

References

[9] Available at: http://stp.lingfil.uu.se/~mojgan/UPC.html