Farsi lexical analysis and stop word list

M.R. Davarpanah
Faculty of Education and Psychology, Ferdowsi University of Mashhad, Mashhad, Iran
M. Sanji
Imam Reza University, Mashhad, Iran, and
M. Aramideh
Mashhad Education Organization, Mashhad, Iran

Abstract
Purpose – The purpose of this article is to present an aggregated methodology for construction of the stop word list in Farsi language and generate a generic Farsi stop word list.
Design/methodology/approach – The stop word list is extracted based on: syntactic classes, domain dependent, corpus statistic and expert judgments. Some of the main challenges that arise in the Farsi automatic text processing are outlined as well.
Findings – Results from the techniques are aggregated and a general Farsi stop word list containing 927 words is generated.
Practical implications – The created stop word list can affect the efficiency and effectiveness of retrieval and indexing process in Farsi information retrieval system, moreover, it can play an important role during Farsi text segmentation.
Originality/value – Our stop word extraction algorithm is a promising technique; it could be applied into other languages that they have ambiguities in automatic text segmentation.
Keywords Languages, Information retrieval
Paper type Research paper

Introduction
In information retrieval (IR), a document is traditionally indexed and searched by words. By definition, stop words are very common words that appear in the text frequently but do not carry significant information in terms of information retrieval (Zou et al., 2006). They can affect the retrieval effectiveness because of a very high frequency and tend to diminish the impact of frequency differences among less common words. They can also affect efficiency, resulting in a large amount of unproductive processing (Korfhage, 1997). The removal of stop words can increase the efficiency of the indexing process as stop words can represent 30 to 50 per cent of the tokens in a large text collaboration (Schauble, 1997). Therefore, identifying a list of stop words in order to eliminate them from text processing during indexing as well as before querying is essential to an information retrieval system.

Lots of stop words lists have been developed for English language in the past, which are usually based on frequently statistics of a large corpus (Van Rijjsbergen, 1975; Francis and Kucera, 1982; Fox, 1990). Farsi like any other natural language
contains many stop words that can represent a considerable percentage of the tokens in a text. However, no general and commonly accepted stop word list has been constructed for Farsi language. Research and experimentation in natural language processing (NLP) in Farsi language is fairly new. There are a few studies on Farsi stop words in the literature. Taghva et al. (2003a, b) in Nevada University produced a stop list that embodied variant forms of only 12 verbs (saying that each verb had almost as many as 100 variations given its infinitive, imperative and past tens forms). The list they produced also included non-verbal stop words which mounted to 155 words. It is an introductory part of a research on Farsi searching and display technologies and created list is relatively short and incomplete as mentioned by the authors. Falahati Qadimi Fumani and Ramachandra (2008) examined 30 chemistry scientific articles to find out that all punctuation marks, numbers, and English letter combinations can be included in the stop list or not. The study revealed that the omission of all punctuation marks, numbers, etc. will have a negative effect on recall. Since punctuation marks, particularly “dot” and “hyphen” appear in the structure of content-bearing elements and even appear abundantly in titles and abstracts of scientific articles. With the fast growth of online Farsi documents and the rapid increase of Farsi web sites, constructing a general Farsi stop word list with an applicable generating methodology becomes critical. This paper represents an exercise in generating a stop list for general text based on the domain dependent and domain independent text collections.

Methodology
Stop lists can be domain independent or domain dependent. They can be created using syntactic classes, or using corpus statistics, resulting in a domain dependent approach for well-defined fields (Abu El-Khair, 2006). This study explore a combination of four techniques of syntactic classes, domain dependent, corpus statistics and expert judgments to obtain the benefits of all approaches. Firstly we reviewed the characteristics of Farsi language and the word structure to identify area of problems for creating stop list. In the next stage a number of 63 articles extracted from 12 different high ranking Persian journal titles of Psychology, Education, and Library and Information Science randomly. To collect the data the articles were typed in a Word format. We manually edited the texts and several fatal spelling errors in the texts were corrected. Then the words automatically segmented. A post-segmentation script was used to join the detached morpheme to the stem and separate the adjacent words prior to word analysis. Syntactic function of the obtained words were examined by using Loghatnameh Dehkhoda, Farhang Moin, Farhang Sokhan, the comprehensive dictionaries in Farsi language.

The research also used corpus created by Hamshahri online Persian newspaper in Iran. The corpus contains 190,206 articles covering a range of different topics such as politics, city news, economics, reports, editorials, literature, sciences, sociology, foreign news, sports, etc. The size of the documents varies from short news to rather long articles with the average of 1.8 KB. More important it consists of real text in everyday use of Farsi that implies it has the sampling and representative property (Darrudi et al., 2004). The data collected through above mentioned techniques function as basis for data analysis. During the stages the research also called the expert judgments.
Characteristics of the Farsi language and area of difficulties

The effectiveness and ability of any IR system to perform efficiently in a language other than English depend on the capability of the system to conform to specific language in use, and understanding the characteristics of the language under study is of great importance (Abu El-Khair, 2006). This section presents an overview of the major characteristics of Farsi language and the main challenges encountered in the producing a stop word list for Farsi.

Farsi, also known as Persian, is an Indo-European language. It is the official language of Iran and is spoken in Afghanistan, Tajikistan and parts of Uzbekistan. A distinguishing characteristic of the Farsi language is the right to left orientation in writing. Farsi uses an expanded form of Arabic alphabet, introducing letters like ch, p, zh, and g. The Farsi alphabet consists of 32 characters that are consonants and vowels: ا، ب،پ،س،ش،ص،ض،ط،ظ،ع،غ،ع،ک،ب،گ،ل،م،ن،و،ی. Moreover, there is a symbol (é) that can be added to the character or the word and change the pronunciation of the character or the word, which makes a total of thirty-three characters. Letters are distinguished by one (ten cases), two (three cases), or three dots (five cases) placed above or below the letter. Three long vowels, AA, EE, and OO are also represented by letters. Short vowels for A, E and O have no letters and may be shown by diacritic marks. In addition, diacritic mark احتمالا is used as final overbar stroke to make an adverb form such as احتمالا (probably). In some cases short vowels can change the meaning of a word bases on their position, e.g. "تش" (full), "پ" (feather). Some of these usages, however, are limited to colloquial speech and they are rarely used in written text, which results in different possibilities of word analysis.

Farsi plural are formed more regularly by addition of suffixes: "ها", "ان" and words borrowed from Arabic "ین". The plural might be formed irregularly depending the root and singular form of the word and a complete reformulation of the word (میثاق) (cause), (means, equipments) (میثاق).

Every letter in Farsi is pronounced as a word cannot be used to represent one character except the waw "و" which means "and. This letter comes in its separate format in a large portion of words in Farsi. Acronyms or abbreviations are not common in Farsi; there are some abbreviations in Farsi that are presented with one-letter words. For example, there are four forms for "ه" (الله) (praise and greeting to God/Mohammad and his descendants) in written Farsi; therefore, except for "ه" (and), all one letter words are abbreviations. Farsi is written using the Arabic script, each character may have up to four different presentation forms in writing according to its location in the word or syntactic situation: initial, middle, final and isolated. For example, there are four forms for "غ"; an initial "غ" (cave), a medial "پ" (Moghul), a final "پ" (razor) and an isolated "پ" (garden). Therefore, each letter may have up to four different shape based on connectivity and its occurrence at the word. The initial form indicates that no element is attached to the element from the right; there is no attaching character before it, but there is one followed the character. Characters are in medial form if they have an attached character both before and after them. The final form denotes that the character is at the end of a word. The final forms can therefore be used to mark word boundaries. The isolated form is a standalone character (Taghva et al., 2003a, b). In Farsi, words may be written as connected or separated characters "میتوم" (I am going). Thus, a single word may have got different presentation forms. In addition, some letters do not connect to others at all, "زود" (soon, quickly, early).
In Farsi texts, word boundaries are identified by using space as word delimiters. But, the inconsistent usage of the whitespace in the texts gives rise to problems in detecting word and boundaries. The optional nature of whitespace causes distinct words to appear as single tokens (e.g. میانکردن = to express), also raises issues in the detached morphemes such as رفته‌ای (you have gone). The inflectional morphemes such as می‌کنم, آمده‌ام, دارم, آورده‌ام can appear either as bound to the host, as free affixes separated by final form character, or separated with an intervening space (e.g. کتاب‌ها = books). The discontinuity may occur in the word structure such as نمایشگاه = display (indexer). There exist lexical elements, such as the preposition بـ, the determiner "نـ", the postpositions "بـ" or the relativizer "کهـ" that usually appear a separate words in written text, but may appear attached to the adjacent word (e.g. ستاره‌ی = intensively). Similarly, a number of pronominal or verbal clitic elements may occur on various of speech categories (e.g. روبروست = is facing). The element within a noun phrase are linked by the enclitic particle called ezafe. This morpheme is usually an unwritten vowel (Megerdoomian, 2004). Lack of an ezafe morpheme delimit the boundary of the pronoun and the proper name in written text as علی شریعتی (Ali Shariat). In such cases the word constructs two words. There are also a large number of multi word expression in Farsi. These include phrasal verbs, compound nouns and lexical units (unit-like element) such as پیش از دریافت (therefore), پیامدهای (not based on). The presence of these multi words make the segmentation difficult because of optional whitespace within them (Megerdoomian, 2004). Farsi verbs are modified more extensively than English verbs. Farsi verbs vary form according to tense, person, negation, and mood. Therefore, a given verb may have a score of variation (Taghva et al., 2003a, b). For instance, Figure 1 shows all variations (present participle) of the verb یاد کردن (to go). According to the figure it has 136 variations.

All these complexities together and multi meaning and multi functioning of the words lead to difficulties in recognizing word boundaries and producing invalid words in the word extraction stage and make processing and identification of stop words in Farsi language more difficult than it appears. Reviewing of the Persian linguistic and grammar literature (Safavi, 1981; Meshkatoddini, 2006; Bateni, 2003) revealed that words in Farsi such as other languages have two distinct level of representation: semantic and syntactic representation. Syntactic words have a finite state and semantic ones have an infinite state. As is mentioned in the related literature stop words serve only a syntactic function. Instead, they are used just because of grammar and carry no significant information (Zou et al., 2006). Therefore, possible words that may be considered as stop words should be collected from the different syntactic classes in Farsi in a systematic way to ensure the completeness of the list. From the viewpoint of linguistics, Farsi stop word usually will be those words with the following word categories: adverbs, pronouns, prepositions, determiners, conjunctions, interjection, ordinal numbers, lexicon units, auxiliaries and light verbs.

As Megerdoomian (2004) believes, listing of the unit-like elements (e.g. پیش از دریافت = therefore) can improve the accuracy of Farsi segmentation significantly. Moreover the subparts of these lexicon units may not appear as index or search terms in Farsi information retrieval system. Most verbal constructions in Farsi are formed using a light verb such as کردن (to do), (to hit), (to give). The number of these verbs is limited but their constructions are extremely productive in Farsi. Light verbs in Farsi behave like single lexical verbs and also function as verbs, when they combine
with different non-verbal items (Karami-Doostan, 1997). Farsi light verbs have many variations and each variation is a poor index or search term. However, a large number of these variations because of their infrequent occur failed to make our stop list. Personals pronouns can appear either as free or as clitics. Although these cliticized pronouns have the same surface form, they can have different functions depending on the part of speech or syntactic context that they appear on. In Farsi just free forms of pronouns can be collected as stop words. The use of a stemming algorithm may filter out attached words such as cliticized pronouns.

**Domain-based**

Text processing of 63 articles resulted in a total of 248,552 occurrences of words. Assigning grammatical categories to words in a text is an important and difficult
component of NLP. The syntactic form of these words were identified based on the mentioned dictionaries, personal and expert judgments. The frequency of the words were also calculated. Then the stop list was constructed on the syntactic function and frequency of the words. Since rule-based technique can accurately annotate grammatical function of the words, it is used to analyze and mark up a given texts. Statistical technique (frequently) is then applied to disambiguate the result or to guess the stop words for the list. The resulting list constructed of 1,291 words in the three domains. Further analysis showed that stop words represents about 39 per cent and symbols 7.4 per cent (both 46.47 per cent) of the tokens in the collected texts. The punctuation marks that appeared in the body of the 63 articles are: - # ( ) { } [ ]; ? ’ " .

Furthermore, the chi-square test was used to judge whether word distribution differences between the three domains can be considered statistically significant or not. The differences was among 20 per cent of the infrequent words (p = 0.05). The distribution of word frequency showed that the highest frequent stop words have a quite stable distribution in different documents and domains.

We wanted a large list to keep many useless words out of indexes, but browsing the data by a small group of experts it was found that many important words as potential index or search terms were occurred in the list; also the presence of such a words may have a negative effect on recall of Farsi IR system. Based on the expert recommendation these words (545 words) were removed, and a list of 746 words produced at this stage. The top ten words are: و, در, به, که, از, این, را, این, با, برای (and, at, to, that/which/who, of/from, this, by, with, for)

Corpus-based
In the next stage, Hamshahri corpus statistic were used to create a corpus-based stop list words occurring more than 20,000 times. Preliminary examination of the corpus word frequency statistics and resulting stop list suggested that this was a reasonable heuristic. A total of 758 words and symbols (737 + 21) met this criterion. The top ten words are: برای, و, که, از, این, به, از, که, دارند, تهران (for, and, country, at, Iran, to, of/from, they do, have, Tehran). Analysis of the punctuation marks appeared within the structure of the corpus revealed that these symbols are the same that were illustrated in the previous pages. Browsing the data from the created list it was also obvious that many words including invalid words, different presentation forms as well as words important as index or search terms such as country, Iran, and Tehran among the top words. A proper segmentation of Farsi texts is required before construction of, because the word boundaries are not clear in Farsi texts. Therefore, a manual check was done to remove any content bearing words, which may not be considered conventional stop words, from this list. The removal of these words (315 words) in another decision based on personal judgment, and syntactic rules (word categories) were identified in this study. The resulting list contained 422 words, as second list of stop words. Comparing the two word list indicated that the content of these stop word list vary in a lot from each others, but quite stable for 246 words (Table I). The chi-square test result (P = 0.01) showed that the difference between the two techniques distribution is significant at the 99 per cent level of confidence. This result indicates that for producing a complete and reliable stop word list in Farsi both statistical and
rule-based as well as domain dependent and independent methodologies should be combined.

Aggregation

The two lists were aggregated together to generate the final one. A combination of these two observations redefined the stop words as those words with stable and high frequency in documents. This list generated according to the above mentioned techniques revealed the features of stop words in different manners, which are all quite reasonable. After removing 246 words in common between the two lists, the stop list is brought to its final of 922 words, that should be maximally efficient and effective in filtering the frequently occurring and semantically neutral words in Farsi texts. These words are listed in alphabetical order in the Appendix (Figure A1). It contains verbal and nonverbal stop words in Farsi almost none of which would be potentially good index or search terms. Our experiments are based on Hamshahri corpus and a text collection in psychology, education, and library and information sciences. It is believed that those stop lists produced for a language in general will not be able to function appropriately in specific domains. Therefore, we recommend that stop lists can be prepared for each subject area separately, i.e. for biology, chemistry, physics, etc. However, the lack of common large collection has been a major draw back in Farsi text retrieval experiments.

English, Arabic and Farsi stop word list comparison

The final aggregated list was checked against the stop word list of SMART (Lewis et al., 2004), which is a standard and well known in English language. We find that about 43.4 per cent (i.e. 248 words out of 571 words) of the English stop words have corresponding words in Farsi stop word list. For instance, "and", "at", "to", "that/which/who", "of", "this", "is", "by", "for". The large number of stop words is due to the characteristic of the Farsi language. However, the specialty of Farsi stop word list is the presence of lexical units, lack of clitic personal pronouns and also that some words might have the same meaning, like "of", "from", both of which means "of".

Because of the close relationship between Arabic and Farsi and the presence of a lot of Arabic loan words in Farsi, we also compared the overlap of our Farsi stop word list with a general Arabic stop word list created by Abu El-Khair (2006). The result showed that only 48 (5.20 per cent) words such as mieux, اخیرا, اکثر, etc. are common between the two lists.

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Number of initial words</th>
<th>Number of removed words</th>
<th>Number of stop words</th>
<th>Number of common words</th>
<th>Number of non-common words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain-based</td>
<td>1,291</td>
<td>545</td>
<td>746</td>
<td>246</td>
<td>500</td>
</tr>
<tr>
<td>Corpus-based</td>
<td>737</td>
<td>315</td>
<td>422</td>
<td></td>
<td>176</td>
</tr>
<tr>
<td>Total</td>
<td>2,028</td>
<td>855</td>
<td>1,173</td>
<td>246</td>
<td>676 + 246 = 922</td>
</tr>
</tbody>
</table>

Table 1. Comparison of the domain-based and corpus-based stop word distribution
Adding words for free
Add words to the list according to the following criteria:

- Our list includes the highest occurrence light verbs, each with infinitive and past tense forms. Though a given verbal root may occur frequently, most of its variations occur infrequently. Therefore, add different variation of the light verbs in list.

- Ordinal numbers in Farsi are made with cardinal numbers plus یا/می/من. Add these for that differ from a word already in list.

- Add any single letter that differ from a letter already in list.

- Add any different spelling of the words included in list (سه = in-time/opportune).

- If a word in the list can take the suffixes یا/ه، ات، تر/ترین, then add the suffixed words (e.g. جدتر = more serious).

- All the suffixed words and unit-like element in the list are written without intervening space, add different presentation forms of these words.

- If a word in the list can take clitics pronouns، "من"، "تان"، "نان"، "شان"، "پایش" = for him. then add the suffixed words (e.g. "آن" = for him).

Conclusion
Traditionally stop words are supposed to have included only the most frequent occurring words. Browsing the data from our lists, domain-based and corpus-based, revealed that stop lists have tended to include infrequently occurring words, and have not included many frequently occurring words as also identified by Fox (1992). Therefore, it can be concluded that statistical model may have poor performance in identifying stop list in Farsi. Since automatic word processing is based purely on spelling and space of words, because of different spelling and space character may or may not be occurred, certain ambiguities arise in a computational segmentation and processing of Farsi texts. In fact the same surface form can represent different morphemes. So our Farsi stop word list facilitates the process of the word segmentation in Farsi information retrieval by increasing the accuracy of segmentation. The stop list that we have generated can serve as the basis for stop lists for specialized databases, or as a list for general Farsi literature.

The major conclusions that can be drawn from this exploratory work are:

- Multi analysis method would have a good performance in Farsi language processing.

- All the words under different syntactic classes, except letters, may not be considered as conventional stop words, as some of them are words important as index or search terms.

- The extremely common words are quite stable among different disciplines; the differences are among the infrequent words.

- Stop words represent approximately 39 per cent of the tokens in Farsi texts, therefore, excluding stop words, especially in full text search system, save space about 39 per cent and speed up searches with little effect on the quality of results.
• About 43 per cent of the English stop words have corresponding words in Farsi stop word list, it can be concluded that translating an English stop word list even augmenting it with high frequency words from the corpus may not be suitable for Farsi texts.

Farsi lexical analysis

References


Fox, C. (1990), Lexical Analysis and Stop List, Prentice-Hall, Upper Saddle River, NJ.


Further reading

Figure A1. List of Farsi stop words

(continued)
(continued)
Figure A1.

(continued)
(continued)
Figure A1.

(continued)