

Providing a Persian Language Singular-Stemmer System (RiCeST Stemmer)

J. Mehrad, Ph.D.

President of RiCeST & ISC, I. R. of Iran
email: dean@srlst.com

S. R. Berenjian, M.A.

RiCeST, I. R. of Iran
Corresponding Author: lang@srlst.com

Abstract

This article aims at defining RiCeST Stemmer in Persian language set up in the Regional Information Center for Science and Technology (RiCeST). We applied linguistic knowledge and standard algorithms to extract machine-readable rules. In addition, plural suffixes and exceptions of which compound nouns are a part were applied. Different parts of Singular-stemmer and their functions are described.

Keywords: Singular-Stemmer Software, Information Retrieval, Singular-stemmer Algorithm, Plural Suffixes, RiCeST, RiCeST Stemmer

Introduction

Language is in contact with the outer world from two aspects: speech and writing. When speaking happens, it can be in two forms: speech or writing. If it is speech, the first material will be sounds, and if it is writing, the first materials are conventional visual signs which are used by human societies and belong to outer physical world (Bateny, 1969).

Of course, in human languages, writing does not represent the sound system perfectly and this is a fact that writing is not accurately equivalent to speech (Anderson, 2002). However, writing or it is better to say conventional visual signs have been the only way to record and preserve the sounds for too many years.

In spite of machines like tape recorders, DVDs, CDs, etc. we use writing for data transmitting, information and document storage and retrieval. In order to make tools, models or techniques that would be applicable to information retrieval, it is important to pay attention to linguistic and Natural Language Processing.

Information retrieval decides to retrieve documents that meet the user needs. User need is in accordance with query portal and it includes one or more keywords. Therefore, Information Retrieval occurs according to matching query with keywords (the most important ones) in documents. The decision can be accepted or rejected, or it can be done according to conjecture of relevance ratio_between documents and query.

Unfortunately, words that are found in documents and queries often have morphological variants. So, a pair of words like “Ab” and “Abyari” is not recognized as equal without processes of Natural Language Processing.

In most cases, different forms of words have similar semantics and can be used as an equivalent for applicable goals in information retrieval. So, some stemming algorithms or stemmers are produced to stem the words.

Therefore, keywords in query or documents are displayed in the form of infinitive or stem of the words instead of the main words.

Thus, different forms of a term are combined as unique form.

Also the number of terms needed for displaying a collection of documents reduces. Smaller forms of terms save space and time for processing. Google and Lycos search engines are examples of systems that use stemming algorithms. The Regional Information center for Science and Technology (RiCeST) is going to install RiCeST Stemmer in its search engine in order to manage databases of this center.

Review of Literature

Little has been done in research about information retrieval and Persian language. The most important ones have been during 2004 to 2005 in Tehran University as two workshops. Articles of these workshops are about Persian language and computer (Bi Jan Kan, 2004). These workshops include the collection of summaries of articles and their full texts in different contexts about Persian language potentials in electronic formats. Participants reported their experiences with Persian language and computer.

As a whole, their linguistic knowledge of Persian language was poor and imperfect and it weakens their assessment and their tool's output.

Neshat (2000), in her article on Persian language and script difficulties, believes that singular and plural subjects can not provide a suitable solution to solve these problems in information science (Alizadeh, 2009).

Kazem Taghva who studies the Persian stemmer, believes that stemmer's effect on the big collections is vague and the stemmers must be modified to improve the effects on information retrieval (Mehrad, 2008).

Jalili (2004) introduces Persian Stemmer tools in his article and names some of its applications such as context automatic classification in big archives, indexing the words, searching the texts in search engines and mechanical translation. But it does not introduce any products (Alizadeh, 2009).

Non-Persian Stemmers

The first stemmer was published in 1968 by Julie Beth Lovins (Lovins, 1968). It was extraordinary in its time and influenced subsequent works in this area. Another stemmer

was written by Martin Porter and was published on July, 1980 in the journal of *Program*. Among all stemmers this one influenced other stemmers the most.

Both of these stemmers that were the most important ones were mono- language. They were used only in English language. Moreover, more studies have been done on other languages especially Arabic. Al-Gaphari has pointed it in his article (Al-Gaphari, 2010). Multilingual stemmers are also written and morphological rules apply to one or more languages simultaneously instead of rules that interpret search of terms in only one language.

RICeST Stemmer

Most of the stemmers use the same algorithm like Porter (Porter, 1980) for stemming and tools. For this reason they have the same advantages and disadvantages. And also they have minor differences like differences in lists and the number of rule bases and etc. In most of these tools, lack of linguistic knowledge is quite clear.

Therefore, we applied linguistic knowledge and standard Algorithms with supporting of 10 suffixes and almost 2000 exceptions (also irregular nouns) in order to make RICEST Stemmer.

Singular-Stemmer

In order to stem, a list of ten plural suffixes was provided, that can be used for nouns, adjectives or pronouns. Basically, there are two types of plural suffixes. The suffixes that are specialized for Persian nouns and we added 'ha' and 'an' to singular nouns to make plural nouns such as 'ghazalha', 'dokhtarha', 'bolbolan' and 'madaran' (Moin, 1984).

The other suffixes such as 'at', 'in', 'un' (Khanlari, 2005 & farshidverd, 2003) which are used as plural suffixes in Persian language are derived from Arabic language and they are mostly used to make plural Arabic nouns like Morabbajat, Majhoulat, Kalamat, Sareghin, Taherin, Rohaniyoun, Estedlaliyoun. We should note that the presence of one of these suffixes in the place of plural maker prevents the presence of the other plural suffixes. A suffix which is added to a noun might require adding a letter between them or changing; For example, Chaharpa + an → Chaharpayan, 'y' will be added before 'an'.

It is not a fixed rule because sometimes 'k' will be added before 'an'; for example, nia + an → niakan or mourche + an → mourchegan, 'h' changed to 'g' (Meshkat-Aldin, 2000 & Anvari, 2006). And this is not also a fixed rule because in 'pelekan', 'h' is changed to 'k' (Lazar, 2005) or in 'neveshtejat', 'h' is changed to 'j' or in torshi + at → torshijat, 'j' is added between 'y' and 'a' (Meshkat-Aldin, 2000).

Therefore, because the various forms of adding the suffixes to various nouns can be seen in Persian, the exceptions will be increased. If rule bases will be provided, an algorithm should be prepared to check the nouns from the end and work upside-down.

In the next stage, we first identify a letter, then the next letter and if it is according to the rule bases it will be compared to the third letter.

Because plural suffixes in Persian language are not more than three letters, there is no need to check more than three letters of words.

The plural suffixes and the formation of them can be seen in Table 1.

Table 1

Types of Suffixes on the Basis of the Word's Last Letter Sounds

Plural suffixes	last letter is /ā/	last letter is /e/ (h) Which is not pronounced	last letter is /u/	last letter is /o/ (y)	Some exceptions
-ha	-	-	-	-	-
-an	-Yan (aghayan)	-gan (parandegan)	Wan (hendowan)	An (faransavian)	Kan (niakan-pelekan)
-at	Jat (Talajat)	Jat/at (darajat) (rouznamejat) {(h) will be omitted}	-	Jat (torshijat)	
-in	-	-	-	-	
-oon	-	-	-	-	

In addition to the sound rules of the word's last letter there are other rules which are according to syntactic forms and specify the type of the plural suffixes. We can see these rules in Figure 1.

As it can be seen, if 'ha' and 'an' are added to the end of the singular words, plural nouns will be formed (Anvari & Ahmadi Givi, 2006). So, 'ha' and 'an' are plural suffixes in Persian language. The rest of the suffixes which are common in Persian language are derived from Arabic language.

Sometimes these rules are generalized to the words that are not derived from Arabic language (Lazar, 2005).

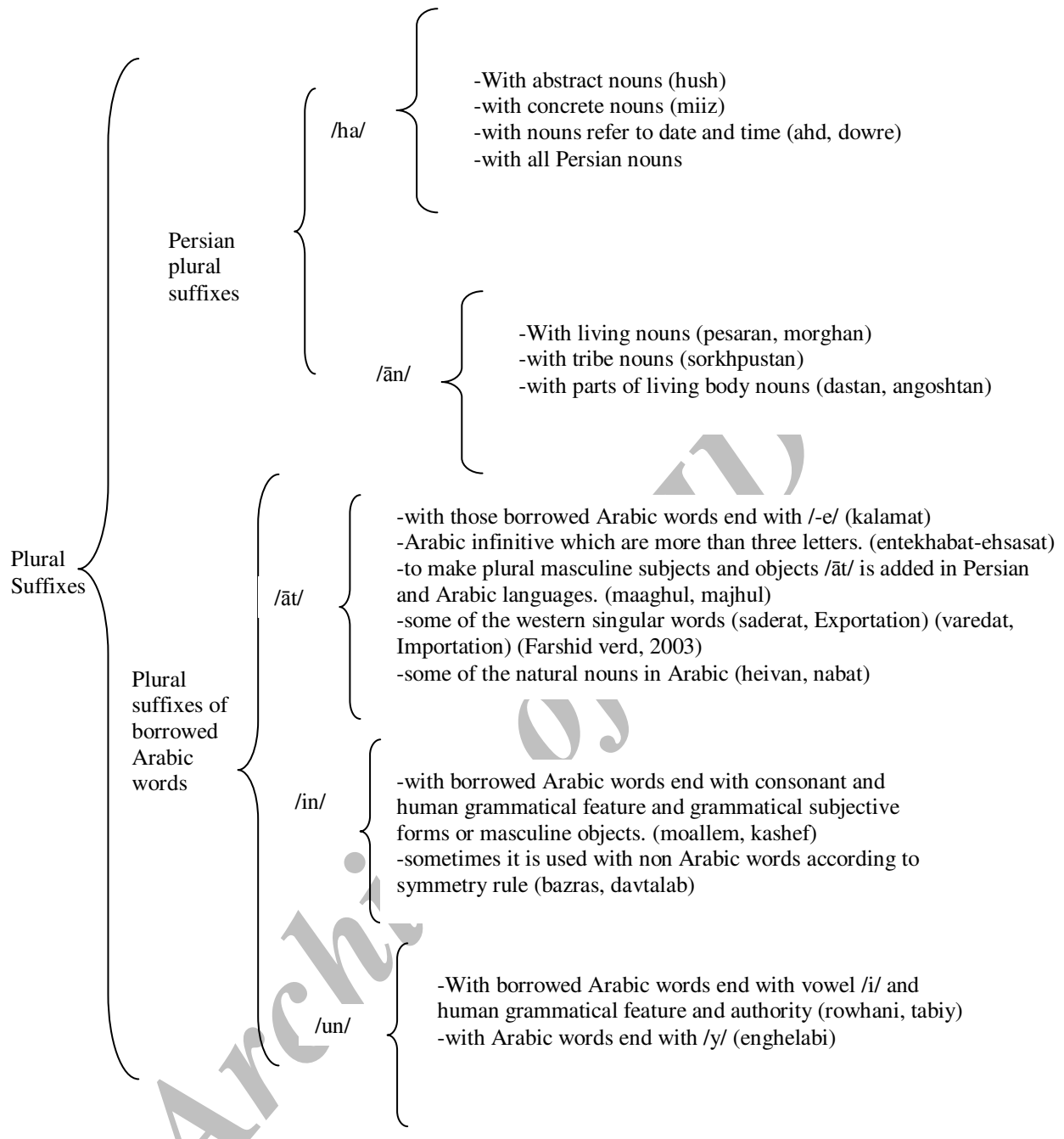


Figure 1. Application of plural suffixes on the basis of syntactic and inflective features

Singular-stemmer system

This system is capable of making singular nouns from plural form. The system consists of these parts:

- A- The distinction between singular nouns and plural nouns
- B- The distinction of plural suffixes
- C- Omission of plural suffixes and presenting the singular nouns
- D- The distinction of the exceptions

- E- The distinction of irregular plural nouns and presenting the singular form
- F- The distinction between plural nouns and singular nouns

More than 11 rules are defined in this system. One of the advantages of this system is that the user can personalize the system based on his need and then evaluate the results. Because all the rules are entered through the numbers, the user can provide a statistical report of his work. In this system, first the user can provide one word at a time and then observe the singular form.

Advantages of S-Stammer system

We can take advantage of this system in many ways, such as

- to reduce different forms of a noun
- to classify the text in the big computerized files automatically
- to conduct research on the root of the words to compress data in information

systems

Different parts of a Singular-Stemmer are provided in Figure 2

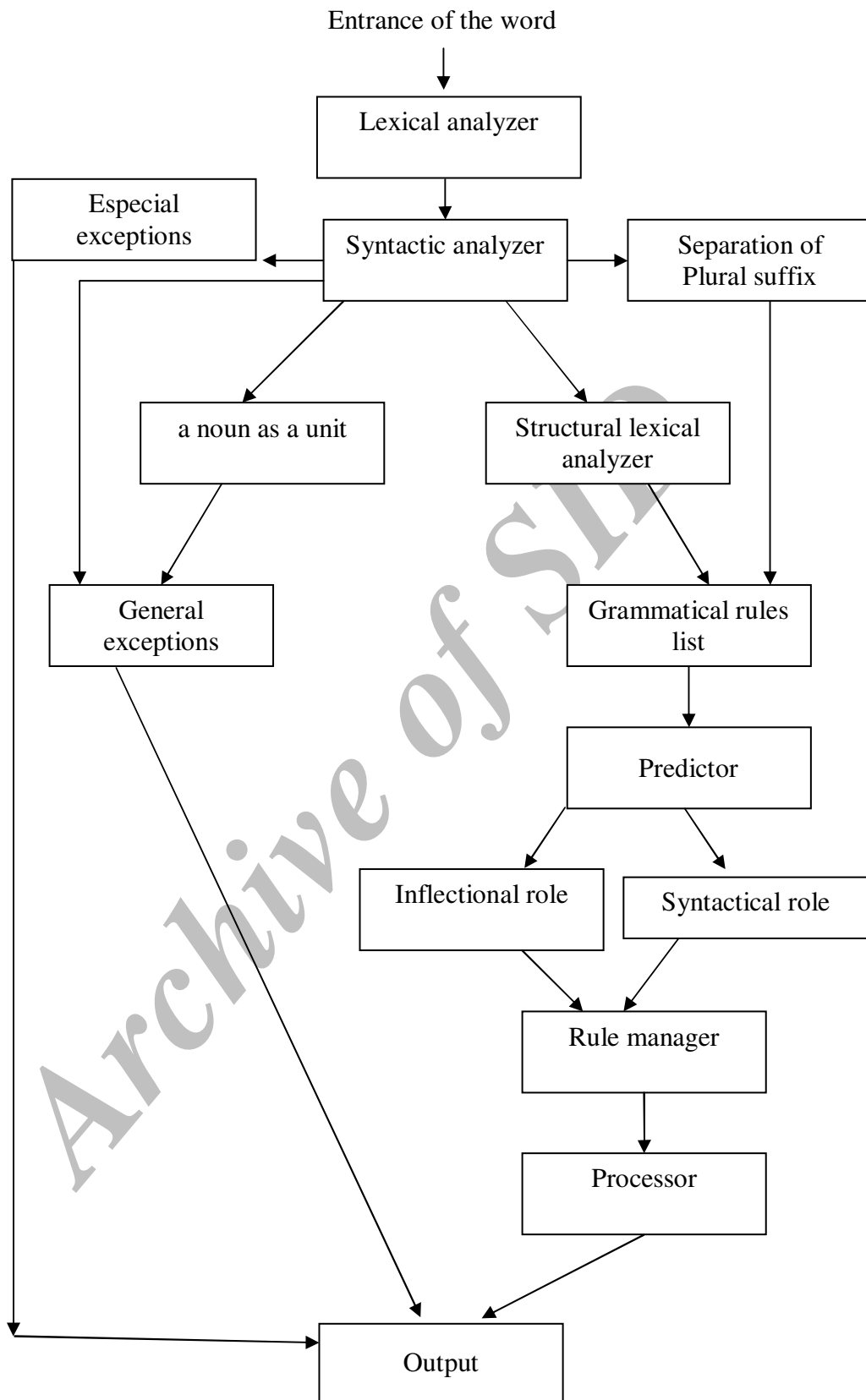


Figure 2. Different parts of singular- stemmer system

The function

After entering each plural noun, lexical, syntactical and structural lexical processing will be applied to each noun.

The processor which has language and pattern-based methods has 4 main parts:

1- Syntactic analyzer: it is a vertical chart analyzer which changes a noun to arboreal list which analyzes the nouns with suffixes.

2- Structural lexical analyzer: whenever the system meets an unknown noun, this analyzer tries to reduce or to stimulate the noun with a pre-determined form.

3- The whole noun as a lexical unit: in case these two above parts would not be accomplished, the whole noun will be taken into consideration as a lexical unit.

4- The forth part, are the especial exceptions that identify the words which are according to the pre-determined words and try to make them singular and finally they move to output.

To extract noun lexical knowledge, first a list of grammatical rules which is in accordance with the current nouns is prepared. Supposing that all the entered nouns are grammatically correct (permissible), the list of grammatical rules allows the predictor to determine the inflectional or syntactic role(s) of unknown nouns.

In this stage, the new noun with its extracted qualities will be sent to grammatical rules to be added to the list if it is needed.

Finally, the plural form of the nouns will be changed to the singular form by Natural Language Processor.

In this processor, all the plural nouns in Persian are divided into twelve groups:

1- The nouns that will be changed to singular form with the omission of 'ha' from the end of them like 'ketabha'

2- The nouns that will be changed to singular form with the omission of 'an' from the end of them like 'koodakan'

3- The nouns that will be changed to singular form with the omission of 'gan' and adding 'h' from the end of them like 'fereshtegan'

4- The nouns that will be changed to singular form with the omission of 'yan' from the end of them like 'danayan'

5- The nouns that will be changed to singular form with the omission of 'wan' from the end of them like 'banowan', 'ahoowan'.

6- The nouns that will be changed to singular form with the omission of 'in' from the end of them like 'motarjemin', 'moalemin'

7- The nouns that will be changed to singular form with the omission of 'un' from the end of them like 'rohaniun', 'alaviun'.

8- The nouns that will be changed to singular form with the omission of 'at' and adding [(e) (h)] from the end of them like 'kalamat', 'shajarat'.

9- The nouns that will be changed to singular form with the omission of 'at' from the end of them like 'pishnahadat', 'maghamat'.

10- The nouns that will be changed to singular form with the omission of 'jat' from the end of them like 'sabzijat', 'torshijat'.

11- The nouns that will be changed to singular form with the omission of 'jat' and adding [(e) (h)] from the end of them like 'neveshtejat', 'resalejat'.

There are also some exceptions that have no rules. Of course, the irregular plural nouns and plural nouns are exceptions; for example, 'ketab', 'kotoob' and 'fazel', 'fozala' and 'galeh', 'lashkar'...

Conclusion

Persian Stemmer software had not been designed in Iran so far. The materials that are written in this case and challenged in various scientific meetings are not based on the software practical observation and are always based on theoretical subjects. This system is installed for the first time in the Regional Information Center for Science and Technology. This study is part of a project and the detailed information is in the project report. This system is used in local network of RICEST.

Before, other stemmers were used and it was a failure for the researchers. So in order to produce the efficient stemmer in Persian language, the local technical knowledge must be produced.

Also, exact scientific investigation of linguistic features of Persian language is the main requisite to meet the desirable goal.

The use of S-stemmer in Persian language will guarantee the desirable efficiency of stemmer.

References

- Al-Gaphari, G.H. & Al-Yadoumi, M. (2010). A method to convert Sana'ani accent to modern standard Arabic. *International Journal of Information science and Management*, 8 (1), 39-49.
- Alizadeh, H. (2009). *Adaptation of librarianship science (PhD dissertation)*. Mashhad Ferdowsi University.
- Anderson, H. R. (2002). *Language and Linguistics*. (Mohammad Reza Bateni, Trans). scientific and cultural publication.
- Anvari, H. & Ahmadi Givi, H. (2006). *Persian Language Grammar (2nd Ed.)*. Tehran: Fatemi Publication.
- Bateni, M. R. (1969). *Description of structure of Persian Language*. Tehran: Amir Kabir Publication.
- Bi Jan Khan, M., Ghasedi, M. E. & Pakdel, M. (2004). *The collection of speeches, reports*

- and abstracts of the projects.* Tehran University and Hushmand processing college.
- Croftand, D. W. & Cruse, A. (2005). *Cognitive Linguistics*. Cambridge: Cambridge University Press.
- Eslami, M. et al. (2004). Computerized changing the text to speech. The First Persian Language Research Workshop and Computer. Tehran University.
- Farshidverd, K. (2003). *Today elaborate Grammar*. Tehran: Sokhan Publication.
- Jalili, S. (2004). Persian words stemmer tools. *The collection of speeches, reports and abstracts of the projects of the first Persian language scientific workshop and computer*. Tehran diploma.
- Kalbasi, I. (1992). *To make derivative words in today Persian language*. Tehran: Studies and Cultural Researches Publications.
- Khanlari, P. (2005). *Persian language grammar*. Tehran: Tous Publications.
- Lazar, J. (2005). *Contemporary Persian grammar* (Bahreini, M. Trans.). Jems Publications.
- Lovins, J.B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, (11), 22-31.
- Mehrad, J. & Naseri, M. (2008). *Natural language processing and information retrieval*. Shiraz: Chapar Publications.
- Meshkat- Aldini, M. (2000). *Persian grammar based on Gaghtari theory*. Mashhad: Mashhad Ferdowsi University Publications.
- Moin, M. (1984). *Singular and plural*. Tehran: Amir Kabir Publications.
- Neshat, N. (2000). Persian writing problems facing information new technology in computerized lists: application and development. In *The articles, application and development of computerized lists in Iran libraries*. Tehran: Information and Constructiveness Scientific Centre.
- Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14 (3), 130- 137.