World Applied Sciences Journal 8 (9): 1078-1085, 2010
ISSN 1818-4952

Automatic Stemming of Some Arabic Words Used in Persian Through Morphological Analysis Without a Dictionary

[1]Ahmad Yoosofan, [2]Ali Rahimi, [3]Mohammad Rastgoo and [4]Mohammad Mahdi Mojiri

[1]Department of Computer, Faculty of Engineering, University of Kashan, Kashan, I.R. Iran
[2]Department of English, Faculty of Humanities, University of Kashan, Kashan, I.R. Iran
[3]Department of Persian, Faculty of Humanities, University of Kashan, Kashan, I.R. Iran

Abstract: Persian language is teeming with Arabic words and there is a need for Iranians to have access to some instrument which helps them differentiate between the Persian and foreign words. One such instrument is stemmer. A good stemmer for Persian must detect and stem these words properly. Such stemmers are by no means free from problems. The basic problem for stemming these words, with respect to Arabic, is their development and the changes they go through. Morphologically, Arabic words have different derivational behavior as compared with those of Persian. Furthermore, some of these words in Persian have specific features which help us distinguish them from Arabic words. To achieve the proper results, we have restricted ourselves to the derivation of some regular triliteral roots. The findings of this research can be utilized in the areas of information retrieval, text categorization, text summarization, automatic detection of phrasal categories, translation studies, natural language processing, etc.

Key words: Computational linguistics · Morphology · Persian stemming · Information Retrieval · Arabic words · Lexicon

INTRODUCTION

In terms of historical linguistics and language family tree, Persian is regarded as an Indo-European language spoken in Iran, Afghanistan, Tajikistan and Uzbekistan. It is derived from the language of the ancient Persian. The different varieties of Persian have official-language status in Iran, Afghanistan and Tajikistan [1]. It is taken for granted that languages are in a state of flux, that is, they influence and are influenced by other languages. In the case of Persian, this language has borrowed a lot of words from Arabic which is another prevalent and influential language in the area. These two languages have proved to be entirely different in certain structures and patterns nonetheless. For instance, Persian and Arabic are originated from two different roots. Arabic is a branch of Afro-Asiatic and Persian is a branch of Indo-European languages [2]. Persian is derived from Indo-Iranian languages, one of the branches of the Indo-European languages. Indo-Iranian split into the Iranian languages and the Indo-Aryan (Indic) languages from which most languages of India are derived. This split is estimated to have taken place around 1500 BC. The major Iranian languages are Persian, Kurdish, Pashto and Baluchi.

Transliteration of Persian and Arabic words has been one problem among others faced by the researchers conducting this study. There are many transliteration standards and suggestions for transliteration and transcription. In this article, we have used transliteration based on [3]. The advantages of this transliteration are as follow: 1. It uses ASCII character code, so preparing words would be simple 2. It could be transformed to correct Persian equivalents and Arabic words automatically. 3. It is an open and free standard mostly accepted by specialists. This standard creates some problem for Persian words such as 1. Rule for tašdīd, that cause some problem in Persian words such as راهها <rah|ha> (roads), توانایی <tawAnAI|I> (competency). On [3] transliteration of latter word is <tawAnA'I> which is incorrect. 2. Transliteration of ع and ء are <`> and <'> respectively, so reading these characters and

Corresponding Author: Ahmad Yoosofan, Department of Computer, Faculty of Engineering, University of Kashan, Kashan, I. R. Iran

differentiating between them is not so easy. Reading this characters in conjunction with other characters, such as f in <Aaf`alA'>, is a difficult task. We should use <> instead of " " for readability. Another good Arabic transliteration is described in [4]. However, this transliteration suffers from the following problems: 1. Unicode encoding is used instead of ASCII encoding. 2. It is designed for Arabic and not for Persian 3. There is no software which automatically converts it to Arabic texts. For more clarification, a complete list of Arabic transliteration is prepared on [5]. However, we have opted for the choice [3]. In appendix, a simple table of this transliteration for Persian is listed.

The influence of Arabic on Persian: Arabs conquered Iran and consequently they affected the way Iranians would speak. Arabic turned out to be the official language of Iranians and the conquest of Iran lasted for four centuries, from the 7th to the 11th AD. Apparently, some intellectual people tried to use Arabic in their discussions, also some writers, poets as well as people in the administration spoke and wrote in Arabic since it was considered as the language of prestige (nowadays English has somehow the same role, it is less common and it is used mostly in academic settings). Moreover, the hierarchical structure of the power relations would necessitate the application of Arabic in court and other arenas of power. Hence, Arabic was the instrument by which people could gain some clout, they simply could not do otherwise. During this period, many Arabic words entered the Persian language. More recently, there has been an Islamic resurgence in Iran since the revolution of 1979 and a considerable number of new Arabic borrowings are being used in Persian writings. These Arabic words have also been added to the lexicon of the language.

Arabic has had an extensive influence on the Persian lexicon, but it has not really affected the structure of the language. Nowadays, analyzing any kind of Persian text, formal or informal, we certainly come across a great number of Arabic terms. It should be pointed out that every software involving Persian processing system confronts these Arabic words and it is supposed to detect and, if necessary, find their relevant stems afterwards. Information retrieval [6], text categorization, text summarization, automatic detection of phrasal categories, translation studies, natural language processing [7], text summarization, error detection and correction, as well as semantic analysis are inextricably tied up with the process of Arabic word detection in Persian texts.

Previous Works

Arabic Stemming: Arabic is a highly inflectional language. This derivational behavior contributes to the preparation of stemmer. However, it has its own specific problems especially in terms of <Ae`rAb> (case signs). Most Arabic stemming methods are based on trial and error. At each step, certain affixes will be removed by resorting to a dictionary and this is done in accordance with the relevant template. Moreover, the remaining words will be searched for in the list of roots (stems). In case it can not be found it in the list, we have to repeat the whole process over and over again. This procedure is carried out as many times as we are able to come up with the intended root [8].

It should be mentioned, however, that there is no one single method for stemming Arabic words. Four different approaches to Arabic stemming can identified: manually constructed dictionaries, algorithmic light stemmers which remove prefixes and suffixes, morphological analyses which attempt to find roots and statistical stemmer which group word variants using clustering techniques [9]. In this study, two kinds of these stemmers have been elaborated in [10] and [11] for those who need further information.

Persian Stemming: Some researches have been conducted in the area of Persian stemming. However, these are much less than the ones carried out on Arabic language. By and large, the extent and depth of the researches on Persian have been far from satisfactory. In [12], [13] the readers can find a list of the studies on different aspect of Persian. Of course, the available Persian stemmers in [14-16] must be added to the previous list.

Arabic Morphology: Morphologically, Arabic is a non-concatenative language. The basic problem with generating Arabic verbal morphology is the large number of variants that must be generated. Verbal stems are based on triliteral or quadriliteral roots (3- or 4-radicals). The stems are formed by a derivational combination of a root morpheme and a vowel melody; the two are arranged according to canonical patterns [8]. This pattern is not necessarily observed in other languages though.

In Semitic languages (a branch of Afro-Asiatic and the ancestor of Arabic) roots interdigitate with patterns to form stems [8]. A number of canonical forms known as measures are used to form verbal stems. Measures are defined as the sequences of consonants and vowels representing word structure. They may also encompass stem derivational affixes.

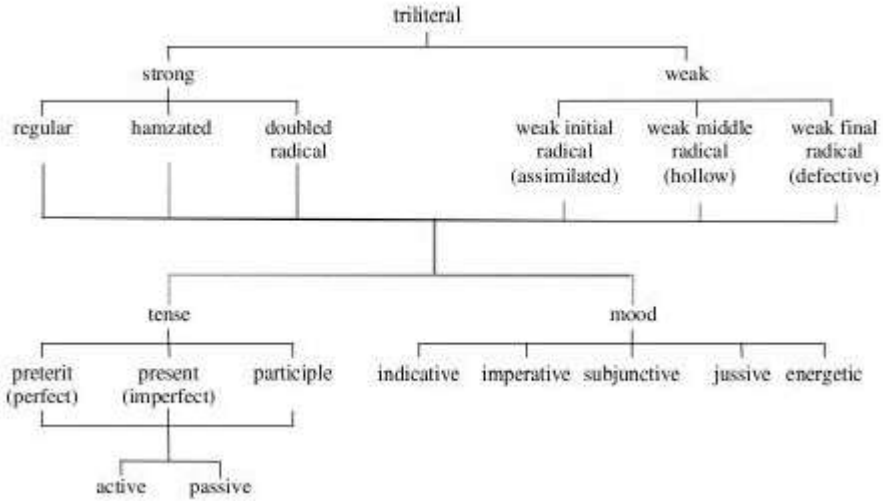

Fig. 1: Classification of Arabic triliteral verbal roots and mood tense system [8]

In addition, perfective (active and passive), imperfective (active and passive) and imperative patterns are usually associated with each measure, which are used to form perfective, imperfective and imperative verbal stems. The perfective verbs signify completed act, while imperfective verbs indicate an unfinished act, which is just starting or it is continuing. Moreover, measures which are intransitive or reflect a state of being do not usually combine with passive voice constructions. The afore-mentioned patterns produce some stems which are used to produce verbs by means of prefixing and/or suffixing inflectional prefixes and/or suffixes [17]. For instance, the Arabic stem darasa (he wrote) is composed of the morpheme d-r-s (notion of writing) and the vowel melody morpheme "a-a". The pattern CVCVC (C=consonant, V=vowel) is utilized here as the criterion based upon which the two morphemes are coordinated. It needs to be mentioned that fifteen triliteral patterns have been observed, at least nine of which have been used prevalently while the other four have turned out be quadriliteral patterns used much less frequently. These patterns are by no means constant and unchangeable, they go through some stem changes in regard with voweling in the two tenses (perfect and imperfect), the two voices (active and passive) and the five moods (indicative, subjunctive, jussive, imperative and energetic). Person, number, gender, tense, mood and the presence of certain root consonants are the determining factors affecting the stem used in the conjugation of the verb. They can change it and produce different conjugations as a result. Triliteral verb roots in Arabic can be classified as shown in Figure 1. Also, a primary distinction is made between weak and strong verbs. Weak verbs have a weak consonant ('w' or 'y') as one or more of their radicals; strong verbs do not have any weak radicals [8].

In this article, the regular strong triliteral Arabic words in Persian are automatically detected and stemmed. Furthermore, the findings of this research can help us find common Arabic words in Persian and prevalent conjugation forms together with the common morphological processes used in Persian. We can also show the effect of some Arabic roots and forms on Persian words and the fact that, under some circumstances, other foreign words borrowed by Persian have tended to ironically and strangely follow Arabic rather than Persian regulations and forms.

Problems Facing the Researcher: The researchers confronted some obstacles in the process of carrying out this research. Stemming of Arabic words in Arabic language is entirely different from that process in Persian. Some Arabic and other foreign words in Persian have undergone changes due to the idiosyncratic characteristics of the Persian language and its rules or its writing system irregularities or congruence (compatibility) with Persian forms as well as general Persian linguistic rules. Anybody embarking on a study of Persian text processing confronts special problems such as:

Lexical changes, 2. Semantic changes, 3. Simultaneous lexical and semantic changes, 4. Spelling changes 5. Changes in grammatical categories of the vocabularies [18].

In [19] many problems of processing Persian text have been discussed. Some other instance of such problems encountered by the researchers will be mentioned in the following parts.

Arabic Tatwīl in the Words: The Arabic character tatwīl < - > (<ka^sIdeH> in Persian) is used in Persian words for fully justifying the words of line. In many office applications this technique is used for justifying lines in a paragraph. Html pages created by those applications include this character and applied on some words, so the same words in these pages may have different characters. This character should be deleted in these words.

Tanwīn: Tanwīn (ـًـٍـٌ) is defined as a neutral "n" plus (o, e, a) attached to the end of some Arabic words [20]. However, sometimes it is attached to the end of Persian and other foreign words in Persian. This is noticed in some Persian words like <^gAnAaN> (sincerely), <zabAnAaN> (verbally), <nA^cArAaN> (inevitably), <telegrAfAaN> (telegraphically) and <telefonAaN> (on the phone). Occasionally, Tanwīn is removed from the words having Tanwīn such as the terms <zAher> (apparently), <_hA.s.seh> (particularly), <mo.sarre.h> (obviously), <.sarI.h> (evidently), <`aIAn> (clearly), <qAa.sed> (intentionally). Sometimes the suffix A replaces the tanwīn as manifested in some words like <Aa.slA> (never), <AabadA> (no way), <.garbA> (western). A lot of adverbs containing tanwīn are constructed at the present time and these words have never been used in this form in Arabic. Words such as <`amIqAaN> (profoundly), <mostaqImAaN> (directly), <mo.tma'yenAaN> (Certainly), <lozwmAaN> (necessarily), <kAmelAaN> (completely), <Aa_hIrAaN> (recently) [18].

Hamza: In Persian, some words are spelled in two different ways and both spellings are equally acceptable. Hamza following the vowel <A> at the end of some Persian words is usually removed. This situation can be seen in word such as <AemlA> must be <AemlA'|> (dictation), <Aen^sA> must be <Aen^sA'|> (essay writing), <Ae^grA> must be <Ae^grA'|> (performance) and <Aa^gzA> must be <Aa^gzA'|> (ingredients) [21].

**Vocalizing (Case Signing, Ae`rAb-Ing):** In Persian writing system, case sign <Ae`rAb> (specifying exactly how each sound is pronounced) is not observed contrary to Arabic. Some writers overuse it and some others do not utilize it at all. These discrepancies in writing practices immensely influence computer processing of Persian texts.

Blending Arabic and Persian Words: Arabic words have been combined with Persian words for a long time and to a very large degree. Hence, separating them from Persian words has turned out to be a difficult task. Persian Plural affixes have been attached to Arabic words and it is absolutely necessary to categorize and recognize the Arabic and Persian words so considerably combined.

Problems and inadequacies of Persian Monolingual Dictionaries: Persian monolingual dictionaries such as <mo`In> and <deh_hodA> dictionaries are out of date and do not contain new terms and expressions which have entered the Persian lexicon through different morphological processes. Most of these new words are related to English technical jargons which have been borrowed by Persian Language. Furthermore, the aim of this study has been the analysis of the words currently used in Persian texts. To make the situation more complicated, some Persian speakers tend to utilize some Arabic words in a different manner. For instance, the Arabic term <AestehlAl> has recently replaced the Persian expression <dIdane mAh> meaning observing the moon at the end of <rama.dAn>. It should be pointed out at the start of this study there was no access to the electronic versions of Persian monolingual dictionaries so the researchers could not analyze the highly frequent Arabic words in Persian. During the first phase of the study, however, <deh_hodA> and <mo`In> Persian dictionaries were obtained and used by the researchers to come up with the necessary statistical information.

Te Marbuta (ة): "Te marbuta" <T> sound ending some Arabic words used in Persian are spelled differently: [22]

- In case it is pronounced at the end of a word, it is written as <t>: <ra.hmat> (bless), <^gahat> (direction), <ne.zArat> (control), <qo.dAt> (lawyers), <morAqebat> (care), <barA'yat> (boycott), <.salAt> (prayers), <me^skAt> (light)
- The letter Arabic <T> is changed into "eH" sound in Persian such as the words <morA^ge`eH> (return), <mokAtebeH> (correspondence), <mo.sA.hebeH> (interview), <mo_dAkereH> (negotiation), <mo.sAdereH> (confiscation), <mo^sA^gereH> (argument) [23].
- Some of the Arabic words infiltrating Persian have split into two words with entirely different meanings. For example, consider the following pairs of sentences:

He/she visited (<morA^g`eH>) the doctor. He/she returned (<morA^ge`at>) from his/her trip.

He/She attended the interview (<mo.sA.hebeH>) for employment. He/She mingled (<mo.sA.hebat>) with his friends. <morA^g`eH> and <morA^ge`at> are from one Arabic word <morA^ga`aT>, also <mo.sA.hebeH> and <mo.sA.hebat> are from Arabic word <mo.sA.habaT>

Table 1: Some Arabic words in Persian

Sample 2				
Type	Transliteration	Meaning	Transliteration	Meaning
Arabic root	D R S		` l m	
Form 1	Dars	Lesson	`elm	Science
Form 2	Dorws	Lessons	`olwm	Sciences
Form 3	Tadris	Teach	ta`lIm	Teaching
Form 4	Not used		ta`lImAt	Teachings
Form 5	Modarres	Teacher	mo`allem	Teacher
Form 6	Modarresin	Teachers	mo`allemIn	Teachers
Form 7	Madreseh	School	Not used	Not used
Form 8	Madares	Schools	Not used	
From 9	Not used		`Alem	Scientist
Form 10	Not used		Ma`lwm	Clear
Form 11	Not used		Ma`lwmAt	Knowledge

Behavior of Arabic words in Persian: This, however is not an innocuously simple process, there are irregularities and aberrations involved. Although a considerable portion of the lexicon is derived from Arabic roots, including the Arabic plural patterns, the morphological processes used to obtain these lexical elements have not been received by Persian and it is not productive in the language. As we notice in the following examples, some of Arabic words together with their relevant roots have entered the Persian language. These words indicate the way the Arabic root system is used to derive nouns by inserting certain vowel patterns in the blank slots in the root template. To illustrate this point, consider some words derived from the Root Form which have entered Persian with different or same meanings in Table 1.

These Arabic words have been borrowed and lexicalized in Persian. Hence, as another example, the Arabic plural form for <ketAb> is <kotob> obtained by the Arabic root derivation system. In Persian, the plural form for the lexical word <ketAb> can be given as in Arabic <kotob>, or it can be obtained by just adding the Persian plural morpheme <ketab>+<hA> = <ketabhA>. Any new Persian word, however, can only be pluralized by the addition of the plural morpheme since the Arabic root system is not a productive process in Persian.

Methods of detecting Arabic words in Persian: We can detect Arabic words in Persian by means of some signs and rules of constructing Arabic words. It may be mentioned, however, that this method is far from flawless.

Tanwīn: Words ending in Tanwīn have got Arabic root with some exceptions.

Words with Te marbuta (round Te in Persian): Arabic words ending in <T> which are used in Persian are usually written in the form of ( ) [22]. Therefore, these words at end of which you can see the letter <T> are necessarily Arabic.

Arabic triliteral measures in Persian: A lot of Persian stems, object nouns, subject nouns and adjectives are made through Arabic triliteral measures. Therefore, we can recognize the Arabic roots which have entered Persian and manifested as different structures. Triliteral used most frequently in Persian are as follows: <Aef `Al>, <taf-`Il>, <mofA`alah>, <tafA`ol>, <tafa``ol>, <Aefte`Al>, <Aenfe`Al> and <Aestef `Al>. Also other derivations from such measures are added to them. Derivations used for detection of Arabic words are as follow:
<mostaf-`al>, <mofta`el>, <monfa`el>, <motofa``el>, <motafA`el>, <mofA`al>, <mofa``al>, <mafA`Il>, <mafA`el>, <Aaf-`alA'>, <fo`lA'>, <fo`wl>, <maf-`alaH>, <maf-`al> and <maf-`Il>. Subject noun and object noun of triliteral are the following: <fA`el> and <maf-`wl>. Most of Arabic words borrowed by Persian are without any signs and often one of the two forms of Arabic triliteral subject noun or object noun has entered Persian. Sometimes they are presented with a small case sign (Ae`rAb). To avoid irregularities, we need to use all forms of vocalizing (complete to no-Ae`rAb)

Arabic plurals in Persian: <At> is the most frequent sign of special Arabic plural such as subject noun, object noun and relative adjective ending in <T>. The following words used in Persian show such a situation:
<ta'asIrAt> (influence), <Ae^stebAhAt> (mistakes), <ta`lImAt> (teachings, instructions), <mo^skelAt> (problems), <moqaddamAt> (introductions or preparations), <ma.h.swlAt> (products)

<Iat> is another Arabic Plural sign:
<`amalIIAt> (operations), <^sar`IIAt> (religious instructions), <rIAzIIAt> (mathematics), ta^grobIIAt (experiences).

Some Arabic plural words in Persian have no singular form; AelahIIAt (theology).
<In> is a sign specific to Arabic words (often subject and object nouns) which are used in Persian: <.hAzerIn> (audience or those present), <qaIebIn> (absentees), <sakenIn> (inhabitants), <mas'ywlIn> (authorities).
<wn> is also a sign indicating Arabic words: <rw.hAnIwn> (clergy), <mellIun> (nationalists), <rawAqIwn> (clergy), <mAddIwn> (materialists), <qe^srIIwn> (partisans) [24].

Implementation: To be able to achieve the goals of the research, we need to carry out a step-by-step organized and systematic implementation process consisting of the following procedure:

Selection of the Words: The words selected for this research have been extracted from a popular daily newspaper in Iran titled "ham^sahrI" (Fellow citizen- from 1996 to 2006). This newspaper is popular newspaper in Iran and has several pages with different subject, so there are many of current words of Persian in this newspaper and we can obtain the popularity of these words. With the help of "httrack" software, this newspaper has been retrieved from the internet. We have designed a program (with python programming language) for extraction of the relevant Persian words with its frequency.

Normalization:

- Blending multi-character letters into one-character letters. For example, three characters for "I" and two characters for "k" merge into one character letters.
- Removing ZWNJ from the end of some words probably because of typos, that is, typographical errors.
- Removing kasreH (kasra in Arabic) at the end of some words and expressions. These kasreH are used to signify adjectives and nouns and certain adverbial clauses in Persian. It may be pointed out that kasreH is not an essential part of such words (not relevant to the words themselves).
- Removing "I" and Tanwīn from the end of words in which <H> (final silent <he>) is their penultimate letter. These two characters function like the previous item.
- Removing Arabic tatwīl which are automatically added to some words in sentences to create balance and harmony in them.

Removing the words identified as Non-Arabic: To decrease the probability of inaccuracy, we'd better ignore the words we have identified as Non-Arabic.

- Eliminating words in which the four Persian letters of (g,^c, p, ^z) are detected.
- Eliminating stop words. These words are extracted from [25].
- Eliminating Persian words and their derivations. These verbs and their derivations are identified by the method referred to in [19].

Identification of Words Which Are Certainly Arabic: Some words are identified as Arabic because of certain characteristics. For example, words having Tanwīn are undoubtedly Arabic (aside from a few exceptions). Also words having the <T> (ـت) are Arabic,as discussed before.

Identification of Other Arabic Words: With the help of different Arabic measures and an analysis of writing form of some words, we can conclude that those words may belong to Arabic stems; we can specify its three probable words afterwards. It is taken for granted that we can not easily claim that the procedure adopted leads to precise Arabic word recognition nonetheless. Hence, a simple model of scaling has been used in this regard. Whenever the numerical scaling exceeds a certain Threshold, we can conclude that the word's root is three letter Arabic and its derivations also relate to the Arabic stem.

The Method of Development for Programming: To harmonize words and one of the Arabic stems, we often make use of finite state technology. Construction, implementation and development of this method is far from easy. Every change in the state machine leads to a change in some codes in the program. Instead of this process, a list of measures were provided and placed in a file specified for stems. It can be concluded that aside from the main letters (root: f, `, l) in stems, the rest of the letters must be the same as word itself. This way, one can easily compare words with their relevant Arabic stems. The textual files of the measures can be modified and edited without the need to change the program. This way the speed of the operation is also immensely increased. Also Persian words from <deh_hodA> and <mo`In> dictionary have been extracted to verify results. Python language is used for implementation of this process.

Table 2:The number of Arabic words detected based on threshold level

Threshold	100	95	90	85	80
Arabic words detected (spotted or identified)	13704	13723	13776	13816	13922

Table 3: Numbers of Arabic stems were found in our collection

Transliteration	No	Transliteration	No	Transliteration	No
Aestef-`Al	130	Aenfe`Al	97	Aefte`Al	240
Taf-`Il	531	mofA`lah	125	tafA`ol	285
Tafa``ol	11	fA`el	3000	maf-`wl	525
Fa`Il	2060	fo`wl	1745	Aaf-`al	7
Fa`lA	103	fa``Al	2191	f-``Aleh	556
Fo``wl	1	fa`Il	1	maf-`Il	314
Maf-`al	1679	maf-`alah	4	mef-`Al	330
Aaf-`Al	1	fa`lA	1737	Aaf-`alA	402
AafA`Il	62	mafA`el	476	mafA`Il	70
Monfa`el	270	mofta`el	246	motafA`el	87

Statistical Analysis: In this research 372249 different words has been extracted from <ham^sahrI> newspaper. 90673 words from among these words are detected as certainly nonArabic words and negative weight is given to them as discussed in "6.3. Removing the words identified as Non-Arabic". 13566 of them are precisely identified as Arabic as discussed in "6.4. Table 2 shows the number of words detected based on the threshold used in the program.

In the Table 3 number of Arabic words is recognized based on triliteral Arabic measures is listed.

CONCLUSIONS

In this article, identification and etymology of some Arabic words in Persian have been undertaken. Some of these words include certain signs and some do not. Those without certain revealing signs have been identified through their conjugation patterns in Arabic since only these words manifest such conjugation patterns. This innovative method opens new horizons for continuation of this kind of research activity and more advanced processing techniques. This study has been the first step in this enterprise and some other aspects must also be taken into consideration. To further strengthen the design of such researches, the investigators must include the rules not discussed in the present study. Comprehensive development of this study has been avoided since the findings need to be corroborated by future studies. It may be mentioned that for the sample under study hallow verbs were not applied and only strong triliteral stems were taken into account. A lot of other details and nuances can also be considered. Other researchers can increase the number of the words under study in their forthcoming researches.

REFERENCES

1. Persian language, Wikipedia, the free encyclopedia, Dec. 2008.
2. Gordon, Raymond G. Jr, (eds.), 2005. Ethnologue: Languages of the World, Dallas, Tex.: SIL International.
3. lagally, K., 2004. ArabTEX Typesetting Arabic and Hebrew.
4. Habash, N., A. Soudi and T. Buckwalter, 2007. On Arabic Transliteration, Arabic Computational Morphology Knowledge-based and Empirical Methods, Netherlands: Springer.
5. Arabic Transliteration, 2009. Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/wiki/Arabic_transliteration.
6. Farhoodi, M., M. Mahmoudi, A.M. Zare Bidok, A. Yari and M. Azadnia, 2009. Query Expansion Using Persian Ontology Derived from Wikipedia. World Applied Sciences J., 7(4): 410-417.
7. Ghaemi, H., 2009. The Effect of Morphological Training on Word Reading and Spelling of Iranian Dyslexic Children. World Applied Sciences J., 7(1): 57-66.
8. Cavalli-sforza, V., A. Soudi and T. Mitamura, 2000. Arabic morphology generation using a concatenative strategy.
9. Larkey, L.S., L. Ballesteros and M.E. Connell, 2002. Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis, Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Tampere, Finland: ACM, pp: 275-282.

10. Al-Sughaiyer and I. Al-Kharashi, 2002. Rule Parser for Arabic Stemmer, Text, Speech and Dialogue.
11. Alserhan, H.M. and A.S. Ayesh, 2006. A Triliteral Word Roots Extraction Using Neural Network for Arabic, Cairo, Egypt: pp: 436-440.
12. Naseh, M., 2007. List of Dissertations on Persian Lanuage and Computer, Tehran university.
13. Shamsfard, M., 2007. Processing Persian Texts: Past Findings and Future Challenges," Tehran University.
14. Yoosofan, 2004. A persian text information retrievel systgem based on latent semantic indexing, Master Thesis, Shiraz.
15. Fuladi, K. and F. Orumchian, 2006. Autonomous Learning, Constructing Persian Words with the Aid of Description of Background and Optimal Length, Tehran University.
16. Hesami Fard, R. and Q. Ghasem Sani, 2006. Designing a Stemming Algorithm for Persian Language, IPM, Tehran.
17. Al-Najem, S.R., 2007. Inheritance-based Approach to Arabic Verbal Root-and-Pattern Morphology, Arabic Computational Morphology Knowledge-based and Empirical Methods, Abdelhadi Soudi, Antal van den Bosch, G¨unter Neumann, ed., Netherlands: Springer.
18. Farshidvar, K., 2004. Modern Comprehensive Grammar, Tehran: Sokhan.
19. Yoosofan, S. Salehi and B. Minaee Bidgoli, 2007. Problems in Stemming Persian Words and a Method for Stemming Persian Simple Verbs, Tehran University.
20. Khayampoor, 2006. Persian Grammar, Tabriz: Sotudeh.
21. Hasanpoor, *et al*., 0000. Educational Approaches, Isfahan: Isfahan Ministry of Education.
22. Adel, H. *et al.,* 2002. A Guidebook to Persian Writing System, Tehran: Persian Language and Literature Farhangestan (Cultural Center).
23. Shariat, M., 1992. Guidebook to Writing, Tehran: Asatir.
24. Vahidian Kamyarm T. and G. Omrani, 2002. Persian Language Grammar (1), Tehran: SAMT.
25. Yoosofan, M., Zolghadri Jahromi and M. Ahmadi, 2005. An Automatic Method for stop word recognition in Persian language, Amirkabir University, Tehran.

Appendix: Transliteration used in this article

Letter	Trans	Letter	Trans	Letter	Trans	Letter	Trans	Letter	Trans	Letter	Trans
ا	A	ب	b	پ	P	ت	t	ث	_t	ج	^g
چ	^c	ح	.h	خ	_h	د	d	ذ	_d	ر	r
ز	z	ژ	^z	س	s	ش	^s	ص	.s	ض	.d
ط	.d	ظ	.z	ع	`	غ	.g	ف	f	ق	q
ک	k	گ	g	ل	l	م	m	ن	n	و	W
ه	h	ی	I	ـَ	a	ـُ	o	ـِ	e	ة	T
ـً	aN	ـٍ	iN	ـٌ	uN	ئ	ý'	أ	'a	ـّ	xx
ه	H	zwnj	\,	(1)	|						