# An Effective Pinglish Filter for Morphological Analyzer of Persian Adjectives

Sude Tavassoli
Department of Computer Engineering,
Islamic Azad University
Lahijan branch, Iran

Samane Tavassoli
Department of Science,
Islamic Azad University
Robat Karim branch, Iran

## ABSTRACT
Natural language processing (NLP) is a field of computer science and linguistics which concerned to the interactions between computers and human (natural) languages. Morphology is the identification, analysis and description of the structure of morphemes and other units of meaning in a language like words, affixes, and parts of speech. In our previous paper a Two-sided morphology analyst of adjectives in Persian were designed. It divides adjective's components into their parts of speech or an adjective can be made. Persian words to English form were converted by us. To solve this problem, in this paper a new filter for converting the Persian words to Pinglish format is designed. Using the adjective as an input of morphological analyzer is so easy.

## Keywords
Mapping Pinglish Filter, Morphological Analyser, Persian Adjective.

## 1. INTRODUCTION
Pinglish is a combination of Persian and English as a term. In Pinglish the English lexical items are nativized and inserted into the framework of Persian morphology and syntax. The history of morphological analysis dates back to the ancient Indian linguist Pāṇini, who formulated the 3,959 rules of Sanskrit morphology in the text Aṣṭādhyāyī by using a Constituency Grammar. Morphology of Language is the identification, analysis and description of the structure of morphemes and other units of meaning in a language like words, affixes, and parts of speech [1, 2, 3, 5, 7]. The form of a word may appear different, morphology have two main branches: inflectional and derivational, that is possible with the help of morphology analysis of each natural way to describe different parts of the language there, and even we can model different dialects [6, 8, 9, 10]. In our previous paper [4] total grammatical adjectives in the Persian language which extracted were about 86 rules and written their Lexicon in Lexc language, and then designed a morphologic analyst in Persian language using the version 8.1.3 Xerox finite state tool.

Lexicons in the XFST converter (Xerox finite-state transducer)[7,13] is implemented so that downward is contain of the input string (word) and upward is included in the base word with its part of speech or vice versa. In this paper, for each part of speech its lexc is written separately, and then has been combined and it would be tried to avoid of additional states and complexity increasing. However, due to increase of adjectives grammatical rules in the Persian language, this machine contains 845 state, 1189 edges and 55.997 paths. Including several sub lexicon and Elementary Lexicon is called Root that indicates starting of network. Root Lexicon is included in adjectives machine 10 sub lexicon, each of which also are divided into several classes. Finally, various adjectives parse to their components and their morphology is done successfully with the help of XfST tool. In Section 2 morphological characteristics studied in the Persian language, in Section 3 defined Lexicons and their Applications and Section 4 is contains the rules of Persian grammar for adjectives, in sections 5 a new Mapping filter for Persian letters to English letters has been designed, the implementation results are at Section 6 and ultimately the conclusions in Section 7 explains the overall discussion.

## 2. MORPHOLOGICAL CHARACTERISTICS IN PERSIAN LANGUAGE
Parts of speech, is one of the most important part of linguistic and is the smallest unit of sentences meaning that brings a concept. In morphology we're looking for every word in sentences, including which parts and which formed part of the letter is. In any natural language vocabularies are classified to different classes. Vocabularies class is include two categories: (1) open class: words Class can be added to their set a new member such as adjectives collection, (2) closed class: Class cannot be added new members and they are fixed, such as Prepositions, conjunctions collection. Overall natural language of the grammar are divided into two groups: language with and without ordering, languages without ordering so to those that are known Ordering rules of a free order in words like Persian language that returns to multi-element form these complex tokens, such as preposition, pronoun, or including the annexes that lexicon category or a separate word component to it that were connected. In Persian language, the smallest component of words conjunctions requires break separately by morphological analysis [6].

## 3. LEXICON DEFINITION AND APPLICATION
In fact, a lexicon is a dictionary that based on all forms of writing should be written. These vocabularies list all form words and phrases that a system should be for recognizing needs. Part of a word may be some types it means that each word can have multi-role [1, 2, 7 12]. To write Lexicon is necessary to achieve the basic grammatical rules for adjective, etc. In the next section to design lexc we have been obtained all Persian grammar rules for adjectives and so it has been designed to Lexicon. We used finite state technology to produce or parse word. This is not logical that we put all Persian words in a database. In order to

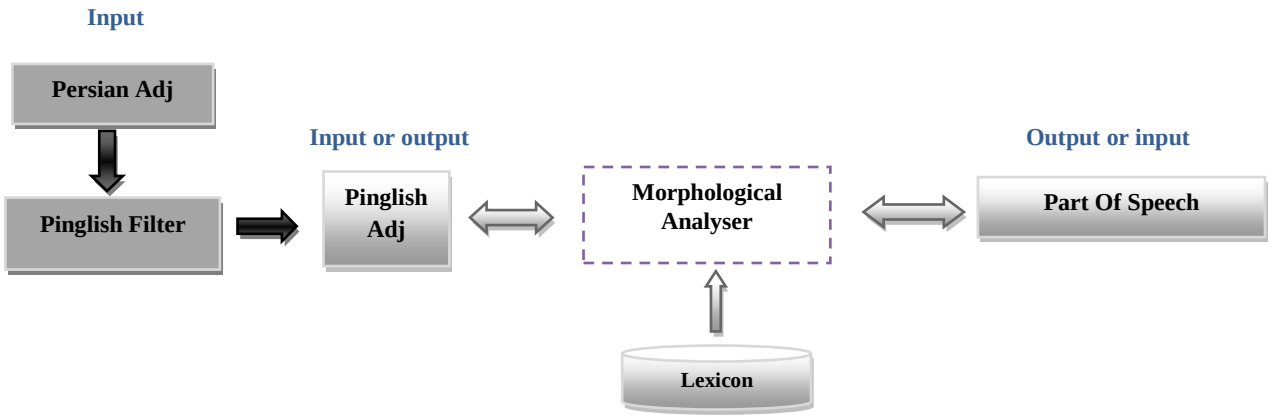


**Figure1. Total Structure of Morphological Analyser[4] and New Pinglish Filter.**

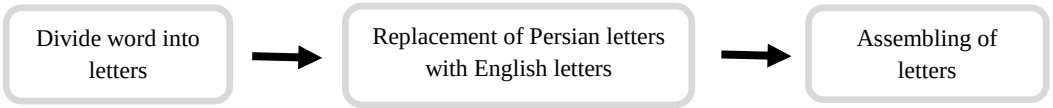


**Figure2. Total structure of new pinglish filter**

use the lexc we should have a tool like xfst. In finite-state system, morphological parsing lexicon is related directly to the content. Thus, if the root form is not listed in the Lexicon component, the morphological parser are created all possible output forms and labeled the word root of the result as "unknown" tag. In this paper, it has been written a large number of adjectives Lexicon and constraint, preposition, conjunction, verb, etc sub lexicon that try by example, putting words in it to implemented output understanding, the other important ability of this system is as well as detect numbers. We can analyze of each natural way to describe different parts of the language by morphology and even model different dialect [9].

## 4. PERSIAN GRAMMAR RULES FOR ADJECTIVES

For writing Lexicon in lexc language, first of all, all grammar rules for the entire Persian adjectives were extracted from references [8,9]. Due to increase of rules just represented a few rules in the paper but the Lexicon diagram given in section 4. Verbal adjective (adjective) is a word that expressed how name intrinsic or spiritual or name successor. Sometimes boundary between noun and adjective is determined, but sometimes is not.

1 - Absolute adjective: adjective that is neither superlative nor comparative such as: bozorg=big (بزرگ).

2 - Subjective adjective: adjective that describes the activity of person such as: presentstem + Participle-Forming Suffixes = zan+ andeh =zanandeh (زن+نده=زننده) (Unfavorable)

3 - Compound adjectives that have more than one component (like derived from adjectives) and simple adjectives that don't have more than one component (solid adjective): noun+noun: sang + del = sangdel (سنگ+ دل =سنگ دل) (cruel).

4 - Counting adjectives: number +number: dou+ hezar= douhezar (دو+هزار=دوهزار) two thousand.

5 - Vague adjectives such as: vague pronoun + noun: ham|n+ kas=ham|nkas (همان+ کس= همان کس) same one, cand+ nafar = cand nafar (چند+نفر = چندنفر) several people

6 - Question adjectives: question pronoun + noun + suffix: ceh +kas+i =cehkasi

(چه +کس+ی =چه کسی) Who.

7 - Combination of two adjectives such as: adjective + Conjunctive +adjective: tar+ va +xoCk =tar va xoCk ( تر+ و + خشک= تر و خشک) dry and wet.

## 5. PINGLISH FILTER DESIGNING

Persian Language alphabet has the 32 letters and 3 low sound signs. For using morphological analyser, we have to transfer the Persian words to Pinglish words. In this stage we use a new filter for mapping the Persian letters to English letters. So the Persian words which can be wrote in the lexicon should be converted into English, the mapping method used for the equivalent Persian to English letters shown as table (1). This filter is implemented in Matlab2008 and it converts the Persian words to Pinglish (Persian word that wrote by English letters) words. Some of mapping words that converted by our Pinglish filter are in below:

سَرد→sard

خوب→xub

بُزُرگ→bozorg

زیبا→ zib|

توانا→ tav|n|

**Table 1. Mapping Persian letters to English letters[5].**

| Farsi letter | English letter | Farsi letter | English letter |
|---|---|---|---|
| آ | A | ض | D |
| ا | \| | ط | T |
| ب | b | ظ | Z |
| پ | p | ع | E |
| ت | t | غ | G |
| ث | V | ف | f |
| ج | j | ق | q |
| چ | c | ک | k |
| ح | H | گ | g |
| خ | x | ل | l |
| د | d | م | m |
| ذ | L | ن | n |
| ر | r | و | u |
| ز | z | ه | h |
| ژ | J | ی | i |
| س | s | ُ | o |
| ش | C | ِ | e |
| ص | S | َ | a |

At first it receive a string (Persian word) as an input, then it replaces the Persian letters and low sound signs with English letters until the end of string. Finally it collects the English letters and produces the Pinglish word as an output (Figure2).

# 6. PREVIOUS WORKS (DESIGN OF MORPHOLOGICAL ANALYSER)

## 6.1. Lexicon designing using generation rules of adjectives

As mentioned each Lexicon as input is given to xfst system and in the output, morphology of adjectives are obtained (Figure 1). Considering much Persian words and because of direct relationship with increasing of network size, for each Lexicon some of words are selected. At first, the grammatical rules of adjectives in Persian language are extracted for designing of lexicons (section4). This two sided morphological Analyzer is shown as figure (1).

## 6.2. Experimental results

To obtain morphology, at first Lexicons that are written in *lexc* format as input are given to Xerox Finite State tool [3,10] as command(1):

Read lexc < file （1）

for analyzing of word in order command (2) an input string (word) is given to XFST it and then word will decomposit to his components and each component will specified what part of speech is:

Apply down *word* （2）

Although generation of an adjective based on a rule, for all components of word with their parts of speech is given to the analyzer as command (3):

Apply up *word*+ part of speech + ….. 3）
（

we can see all adjectives with their components by command (4):

Print lower-words >file （4）

For example, according to rule (2), an adjective such as "bozorgtar"(بزرگتر)(bigger) will breaks to its components with below commands:

xfst[1]: Apply down bozorgtar (5)

bozorg+sefatmotlaq+tar+passwand+suf_suffix

to generate and adjective such as"sangdel"(سنگدل) (cruel) the below commands are used:

Xfst[1]: apply up （6）
sang+esm+del+esm1+esm_suffix +suf_suffix
Sangdel

With command (7) all adjectives can be generated with a total rules:

Xfst[1]: Print upper-words > file （7）

Examples of Adjectives that are obtained using total rules are there:

(weak) ناتوان = n|tav|n
(nice) زیبا =zib|
(fat) چاق = c|q

# 7. CONCLUSION

For using morphological analyzer, we have to transfer the Persian words to Pinglish words. Pinglish is a combination of English and Persian. In Pinglish the English lexical items are notarized and inserted into the framework of Persian morphology and syntax. In this paper we designed a new mapping filter the Persian words to Pinglish words. In our previous paper, at first total rules of grammatical adjectives in the Persian language about 86 were extracted and their Lexicon in Lexc format were written, then a morphological analyser in Persian language was designed using the version 8.1.3 Xerox finite state tool. the lexicons in the XFST analyser (Xerox finite-state transducer) were implemented so that downward was contain of the input string (word) and upward was included in the base word with its part of speech or vice versa. In this paper, for each part of speech its lexc was written separately, and then was combined and tried to avoid of additional states and complexity increasing. However, due to more adjectives of grammatical rules in the Persian language, the lexicons have 845 states, 1189 edges and 55,997 paths. Each lexicon have several sub-lexicon and primary Lexicon is called “Root” that indicates the start of network. Root Lexicon is including of adjectives machines 10 sub-lexicons, that which also are divided into several classes. Finally, various adjectives break into their components and their morphology is done successfully by using XFST tool. According to Experimental Results, each Persian word converts to Pinglish word easily for using of morphological analyzer. Each word with its part of speech is obtained or a word is generated, makes understanding and using this Analyzer easier.

## 8. REFERENCES

[1] Aronoff. M, 2009. "Morphology: an interview with Mark Aronoff". ReVEL, v. 7, n. 12, ISSN 1678-8931.

[2] Haspelmath. M. "Understanding morphology", London: Arnold (co-published by Oxford University Press). ISBN 0-340-76025-7 (hb); ISBN 0340760265 (pbk), 2002.

[3] Laurie. B, "Introducing linguistic morphology (2nd ed.). Washington", D.C.: Georgetown University Press. ISBN 0-87840-343-4, 2003.

[4] Tavassoli. S, Alipour. S, 2010. "A Morphological Analyser For Persian Adjectives and Nouns",in proceedings of IEEE 2010 3۲rd International Conference on Advanced Computer Theory and Engineering ICACTE 2010 (ICACTE2010), volume 5, pages 437-440, Chengdu,China, 22Aug.

[5] Megerdoomian. M, 2006. "Extending a Persian Morphological Analyzer to Blogs", University of Maryland, College Park.

[6] Dehdari. J, 2005."A link Grammar Parser for Persian". Talk presented at the First International Conference on Aspects of Iranian Linguistics, Leipzig, Germany.

[7] Beesley K. R. and Karttunen .L, 2003. "Finite-State Morphology: Xerox Tools and Tecnique s ,CSLI Publications, Palo Alto.

[8] De Gispert. A, Marin. J.B, 2008. "On the impact of morphology in English to Spanish statistical MT", Speech Communication 50, 1034–1046.

[9] Malenica. M, Sˇmuc. T, Sˇnajder. J, B. Dalbelo Basˇic, 2008. "Language morphology offset: Text classification on a Croatian–English parallel corpus", Information Processing and Management 44, 325–339.

[10] Arzhang. Gh, 2005. "dasture zabane farsi emruz" book, fourth Edition.

[11] Shariat. M. javad, 2004."dasture zabane farsi" book, fourth edition.

[12] Megerdoormian. K, April 2000. "Persian Computational Morphology: A Unification-Based Approach", Memoranda in Computer and Cognitive ScienceMCCS-00-320.

[13] Xerox Finite-State Tool at: http://www.cis.upenn.edu/~cis639/docs/xfst.html