A Morphological Analyzer for Persian Adjectives and Nouns

Sude Tavassoli
Islamic Azad University
Lahidjan branch, Iran
sudetavassoli@gmail.com

Sara Alipour
Islamic Azad University
Qazvin branch, Iran
s.alipour@qiau.ac.ir

Abstract- **Natural language processing is a sub-branch of artificial intelligence in which a natural language that is used for communication between humans, is converted to an artificial form. The meaning of morphology is of what components form a word and how these components are put together and create the word. First, in this paper we extract the grammatical rules of nouns and adjectives in Persian (Farsi) language, which are about 86 and 113 rules respectively. Then, we modeled their lexicons in Lexc language and design a Two-sided morphology analyzer of nouns and adjectives in Persian language, using Xerox Finite State Technology such that given an input (adjective or noun), the analyzer breaks it to its components or given the components with their parts of speech, the analyzer generates an adjective or a noun.**

Keywords - Morphology, Adjective, Noun, Persian Language, Lexicon, Part of speech.

I. INTRODUCTION

Language is a set of strings, which are formed of alphabet letters. As we mentioned earlier, Natural language processing is a sub-branch of artificial intelligence in which the natural language which is used for humans communication, is simulated in an artificial format. Morphology is defined as the study of analyzing what components form a word and how these components together can create that word, i.e., how to put words together is called the language grammar [1,3]. The form of a word may appear differently and morphology itself has two main branches, namely, inflectional and derivational. With the help of morphological analysis, describing the different parts of language, and even modeling different dialects are possible [2,4,5,6]. In this paper, total grammatical rules of adjectives and nouns are extracted in the Persian, which are respectively about 86 and 113 rules and their Lexicons in Lexc language are written. Finally, a morphologic analyzer is designed using the version of 8.1.3 Xerox finite state tool. Lexicons in the XFST converter (Xerox finite-state transducer)[3,10] are implemented such that downward is contained of the input string (word) and upward is comprised of the base word with its part of speech, or conversely. In this paper, for each part of speech, its lexc is written separately, and then is combined to avoid having additional states and complexity increasing. However, due to the high number of adjectives and nouns, this machine contains 845, 832 state, 1189, 1170 edges and 55.997, 55.982 paths respectively. It includes several sublexicon and Elementary Lexicons and Root, which indicates the starting of network. Root Lexicons include for adjectives and nouns machines, 10,12 sublexicon respectively, each of which also are divided into several classes). Finally, various adjectives are parsed to their components and their morphology is preformed successfully using XfST tool. In Section II morphological characteristics are studied in Persian language, Section III defines Lexicons and their applications and Section IV contains the rules of Persian grammar for adjectives and nouns. In Section V, a Mapping for Persian (farsi) letters to English letters has been designed, the implementation results are in Section VI and finally the conclusion of the results are discussed.

II. MORPHOLOGICAL CHARACTERISTICS IN PERSIAN LANGUAGE

Parts of speech, is one of the most significant part of linguistic and also is the smallest unit of sentences, where its meaning clarifies a concept for us. In morphology we are looking for every word in sentences including its parts and the part of speech it has. In any natural language, vocabularies are classified in to different classes: (1) open class: a new member can be added to their set such as adjectives, nouns collection, (2) closed class: new members cannot be added and they are fixed, such as Prepositions, conjunctions. In overall, grammars are divided into two groups: language with and without ordering. Those that are known as without ordering, rules have a flexible order in words like Persian language, which leads to multi-element form of complex tokens such as preposition, pronoun, which in the lexicon can be included as a separate word component. In Persian language, the smallest component of words such as conjunctions require break separately by morphological analyzer [2].

III. LEXICON DEFINITION AND APPLICATION

In fact, a lexicon is a dictionary that is based on all forms of writing. The vocabularies list and all forms of words and phrases that a system needs for recognizing. Part of a word may have some different types, it means that each word can have multiple roles [1, 3, 9]. To write Lexicon, it is necessary to achieve the basic grammatical rules for adjectives, nouns, etc. In the next section, in order to design lexc, we have obtained all Persian grammar rules for adjectives and nouns to design the Lexicon. We used finite state technology to produce or parse words. This is not logical that we put all Farsi words in a database. In order to be able to use the lexc, we need to use 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 Lexicon as a component, the morphological parser creates all possible output forms and labels the root of the result by an "unknown" tag.

In this paper, a large number of adjectives and nouns, preposition, conjunction, verb, etc. has been written in sub lexicon, but due to space limitations, only the lexicon of adjectives are shown in Figure 2. The other important ability of this system is detecting numbers. With the help of morphology we can analyze each way to describe different parts of the language and even different dialects can be modeled [6].

IV. PERSIAN GRAMMAR RULES FOR ADJECTIVES AND NOUNS

For writing Lexicon in Lexc language, first, all grammar rules for the entire Persian adjectives and nouns were extracted from two sources [7, 8]. Due to the high number of rules, just a few rules in the paper are represented in the Lexicon diagram. Sometimes boundary between noun and adjective is determined, but sometimes is not. Verbal adjective (adjective) is a word that expresses intrinsic, spiritual or successor of a name.

1 - Absolute adjective: adjective that is not superlative nor comparative such as: bozorg=big (بزرگ).

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

3 - Compound adjectives that have more than one component (e.g., derived from adjectives) and simple adjectives that do not 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 – Derived nouns : prefix +present stem + pastfix =xoud +x|h +i =xoudx|hi (خود +خواه+ ی= خودخواهی) Egoism.

8– Compound nouns: direct object + presentstem : guC +m|l =guCm|l (گوش +مال =گوشمال) punish.

9 –infinitive nouns: past stem + pastfix : Afarin+ eC= AfarineC (آفرین +ش =آفرینش)
Creation.

10 – Indefinite nouns: simple noun + ezafe: pesar +i = pesari (پسر + ی= پسری) a boy

11 – Instrumentation nouns: present stem + ezafe : t|b + eh = t|beh (تاب + ه =تابه) pan.

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

V. MAPPING FARSI LETTERS TO ENGLISH LETTERS

In order to be able to write Farsi (Persian) words in the lexicon, they should be converted into English, i.e., the mapping used for the equivalent Farsi to English letters is shown in Table 1[2].

TABLE I. MAPPING FARSI LETTER TO ENGLISH LETTER [2].

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

VI. DESIGN OF MORPHOLOGICAL ANALYSER

A. *Lexicon designing using generated rules of nouns and adjectives*

As mentioned, each Lexicon is given as an input to xfst system and in the output, morphology of nouns and adjectives are obtained (Figure 1). Considering the high number of Persian words and because of direct relationship with increasing of network size, for each Lexicon some of words are selected. The grammatical rules of nouns and adjectives are extracted for designing of lexicons. This two sided morphological Analyzer is shown as Figure 1. The results are shown below.

B. *Experimental results*

To obtain the morphology, at first Lexicons that are written in *lexc* format as inputs 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 and then the word will be decomposed to its components and each component will be specified what part of speech is :

```
    Apply  down  word                         (2)
```

Although generation of a noun or adjective based on a rule, for all components of word with their parts of speech is given to the analyzer using 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) and a noun like “raft|r”(رفتار) (behavior) will break to their components with commands below:

```
   xfst[1]: Apply down bozorgtar                  (5)

   bozorg+sefatmotlaq+tar+passwand+suf_s
   uffix

   xfst[1]: Apply down raft|r

       Raft+ bonemazi+|r+passvand+sufsuffix
```

In order to generate an adjective such as ”sangdel"(سنگدل) (cruel) or a noun such as “nam|ieCn|meh” (نمایش نامه) (drama) the below commands are used :

```
     Xfst[1]: apply up                           (6)
     sang+esm+del+esm1+esm_suffix
     +suf_suffix
     Sangdel
     Xfst[1]: apply up
     nam|ieC+bonemazi+n|meh+pasvaj+
     sufsuffix :nam|ieCn|meh
```

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

```
    Xfst[1]: Print upper-words  > file            (7)
```

Examples of Adjectives and nouns that are obtained using total rules are here:

(weak) نـاتـوان = n|tav|n
(Spurious) مـن درآوردی =mandarAvardi
(fat) چـاق وچـلـه =c|quceleh
(sweeper) رُفـتـگـر =roftgar
(punish) گـوشـمـال= guCm|l
(a boy) پـسـرک = Pesarak

CONCLUSION

In this paper, at first total rules of grammatical adjectives and nouns in the Persian language were extracted respectively about 86 and 113 and their Lexicon in Lexc format are written, then an morphological analyzer in Persian is designed using the version of 8.1.3 Xerox finite state tool. The lexicons in the XFST analyzer (Xerox finite-state transducer) are implemented so that downward is comprised of the input string (word) and upward is of the base word with its part of speech or conversely. In this paper, for each part of speech, its lexc is written separately, and then is combined. We tried to avoid additional states and complexity increasing. However, due to many adjectives and nouns grammatical rules in the Persian language, the lexicons have 845, 832 states, 1189, 1170 edges and 55,997, 55,982 paths respectively. Each lexicon have several sub-lexicon and primary Lexicon is called “Root” that indicates the start of network. Root Lexicon is of adjectives and nouns machines for 10, 12 sub-lexicon respectively, which also are divided into several classes. Finally, various adjectives and nouns are divided to their components and their morphology is obtained successfully by using XFST tool. According to Experimental Results, each word with its part of speech can be obtained or a word can be generated. This makes understanding and using of the analyzer easier.

REFRENCES

[1] Karine Megerdoomian, “Extending a Persian Morphological Analyzer to Blogs”, University of Maryland, College Park, 2006.

[2] Jon Dehdari. “A link Grammar Parser for Persian”. Talk presented at the First International Conference on Aspects of Iranian Linguistics, Leipzig, Germany, 2005.

[3] Kenneth R. Beesley and Lauri Karttunen. “Finite-State Morphology: Xerox Tools and Tecnique s’, CSLI Publications, Palo Alto, 2003.

[4] A. de Gispert, J.B. Marin, “On the impact of morphology in English to Spanish statistical MT”, Speech Communication 50, 1034–1046, 2008.

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

[6] Chet Creider, Richard Hudsonb,“Inflectional morphology in Word Grammar”, Lingua 107, 163-187, 1999.

[7] Gholamreza Arzhang, ”dasture zabane farsi emruz” book, fourth Edition,2005.

[8] mohammad javad shariat,”dasture zabane farsi” book, fourth edition , 2004.

[9] karine Megerdoormian, “Persian Computational Morphology: A Unification-Based Approach”, Memoranda in Computer and Cognitive ScienceMCCS-00-320, April 2000.

[10] http://www.cis.upenn.edu/~cis639/docs/xfst.html

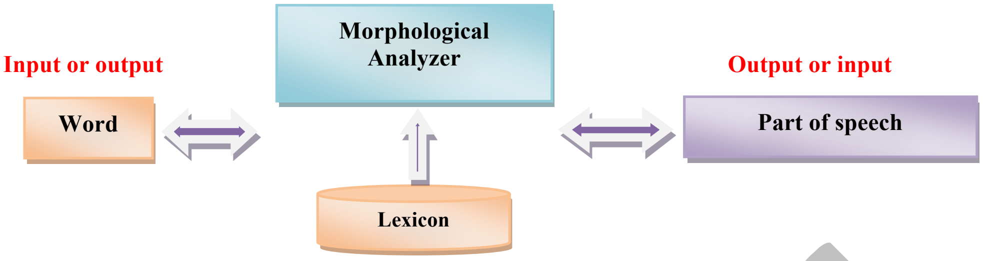

Figure1. The Structure of Proposed Morphological Analyzer.

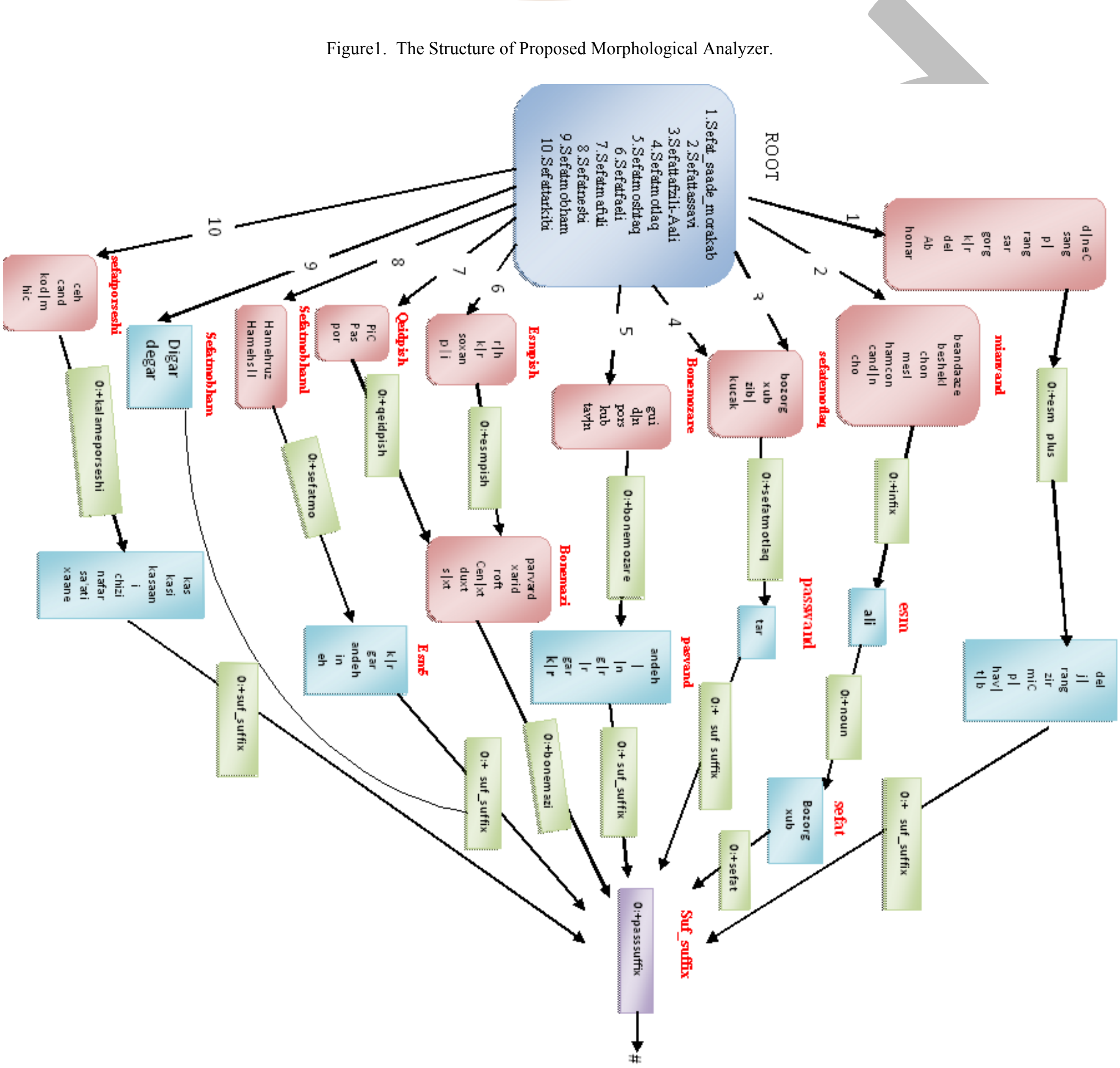

Figure 2. Lexicons for Persian adjectives.