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 write 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.

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:

n|tav|n = ناتوان (weak)
mandarAvardi= من درآوردی (Spurious)
c|quceleh= چاق وچله (fat)
roftgar= رُفتگر (sweeper)
guCm|l= گوشمال (punish)
Pesarak = پسرک (a boy)

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. S˘muc, J. S˘najder, B. Dalbelo Bas˘ic "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

Figure 2. Lexicons for Persian adjectives.