International Journal of Information Science and Management
Vol. 17, No. 2, 2019, 119-134

Word Sense Disambiguation Focusing on POS Tag Disambiguation in Persian: A Rule-based Approach

Elham Alayiaboozar
Assistant Prof. Iranian Research Institute for Information Science and Technology (IranDoc)
Corresponding Author: elham_alaee2000@yahoo.com

Amirsaeid Moloodi
Assistant Prof. Department of Foreign Languages & Linguistics,
Shiraz University
amirsaeid.moloodi@gmail.com

Manouchehr Kouhestani
Assistant Prof. Department of Foreign Languages & Linguistics, Shiraz University
manouchehr.kouhestani@gmail.com

Abstract

The present study deals with ambiguity at word level focusing on homographs. In different languages, homographs may cause ambiguity in text processing. In Persian, the number of homographs is high due to its orthographic structure as well as its complex derivational and inflectional morphology. In this study, a broad list of homographs was extracted from some Persian corpora first. The list indicates that the number of homographs in Persian corpora is high and homographs with high frequency are those that occur as a result of the identical orthographic representation of some inflectional and derivational morphemes. Based on the list, the most frequent homographs are nouns and adjectives ending in <ی> /i/. POS tag disambiguation of such homographs would make word sense disambiguation easier and lead to better text processing. In this study, a list of noun and adjective homographs ending in <ی> is extracted in order to decide their correct POS tag. The result was studied to extract context-sensitive rules for allocating the right POS tag to the homograph in syntactic structures. The accuracy of rules was checked, and the result showed that the accuracy of most rules is high which proves most rules are true.

Keywords: Homographs, POS tagging, POS Disambiguation, Noun and Adjective Homographs Ending in <ی>, Context-sensitive Rules.

Introduction

Ambiguity refers to a situation where a word or sentence can have more than one meaning. A sentence is considered ambiguous if it contains ambiguous word(s). It is worth mentioning that intonation and punctuation changes may also lead to ambiguity; however, only the ambiguity at the word level is going to be studied in the present paper. Practically, any sentence that has been classified as ambiguous, usually has multiple interpretations, but just one of them is considered as the correct one (Abed, Tiun, & Omar, 2015). Ambiguity is one of the main challenges faced in the analysis of natural languages using computers. There are different kinds of ambiguity at word level or sentence level with regard to the word internal structure (which is called morphological ambiguity). An English example includes the

English verb form <look> with no affix: it can either be the infinitive or a first or second person singular/plural verb form, but as soon as the word immediately preceding <look> is taken into consideration, the ambiguity can be resolved in most cases. The same holds true for many other languages including Persian. For example the word <شکست>/ʃ ekast/ in Persian can be either a noun (which means "failure" or "defeat") or a verb (past tense which means "it broke"). Another kind of morphological ambiguity occurs when affixes are added to the root/stem for inflectional or derivational reasons. For example the Persian word <جوانی> /ʤavɒ ni/ may be analyzed as follows: <جوان> /ʤavɒ n/ (young) + <ی>/i/ (second person singular morpheme) = you are young, <جوان> /ʤavɒ n/ (young) + <ی> /i/ (noun maker suffix) = youth, or <جوان>/ʤavɒ n/ (young) + <ی> /i/ (indefinite morpheme) = a young person. There is another kind of ambiguity called lexical ambiguity at the word level which occurs when a single word is associated with multiple senses which itself is traditionally subdivided into polysemy and homonymy (Gaustad, 2004).

As mentioned before, another kind of ambiguity is found at the sentence level, known as syntactic ambiguity. A classic example is the case of PP attachment ambiguity which is found in many languages including English. The sentence "the man saw the girl with the telescope" is ambiguous as it may either mean "the man had the telescope and was using it to see the girl" or "the girl was carrying the telescope."

The present study deals with ambiguity at the word level (the so-called morphological ambiguity) focusing on homographs. Homographs are words whose orthographic forms (spelling) are the same, but their meanings (and sometimes, pronunciations) are different (Merriam Webster dictionary). In various languages, homographs may cause ambiguity in text processing. It seems that English has a "shallow" orthography, there usually exists one pronunciation per spelling (Gottlob, Goldinger, Ston, & Orden, 1999). As a result, there are fewer than 20 common homographs in English. However, in Persian, the number of homographs is high due to its orthographic structure as well as complex derivational and inflectional morphology. In the Persian writing system, short vowels are usually absent and just a few graphemes in a few words are used to represent short vowels, like <ه> which could stand for the short vowels /e/ or /a/ in a few words like <به> /be/ (to), <نه > /na/ (no). The absence of short vowels in the Persian writing system leads to ambiguity in text processing. For example the orthographic form <مردم> has three phonological representation at least: /mardam/ (I am a man), /mordam/ (I died), and /mardom/ (people) (Megerdoomian, 2000). Some other kinds of complexity in the Persian writing system are caused by diacritics which are mostly considered as bound graphemes. The absence of some of these diacritics in different texts may create some homographs, for example the absence of the diacritic referred to as "Tashdid" (Geminatioan) <ـّـ> leads to homographs like <سرّ> /serr/ (secret) versus <سر> /sar/ (head) (Alayiaboozar and Bijankhan, 2013). Part Of Speech (POS) disambiguation of such homographs would make word sense disambiguation easier and lead to better text processing. POS tagging is the ability to computationally determine what POS tag of a word is activated by its use in a particular context (Zeroual, Lakhouaja & Belahbib, 2017). Actually a Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns a part of speech to each word (and other tokens), such as noun, verb, adjective, etc. The present study introduces a way for POS disambiguation of the most frequent noun and adjective homographs. In this study, different classifications of Persian homographs are presented, then the frequency of homographs is studied in three Persian corpora including the

Persian written corpus or Peykare, also known as Bijankhan corpus (Bijankhan, Sheykhzadegan, Bahrani & Ghayoomi, 2011), the Farsi linguistic database, also known as paygah-e dadegan-e zaban-e Farsi (Assi, 1997), and the Persian syntactic dependency Treebank (Rasooli, Kouhestani & Moloodi, 2013). Then, the most frequent ones are studied in the syntactic context using the knowledge of neighboring words with regard to a history of 10 windows (considering 10 left context words and 10 right context words) in order to decide on the right POS tag of the homograph based on the structure of the sentence. Finally, the result is studied to extract context-sensitive rules for allocating the right POS tag to the homograph in syntactic structures and the accuracy of rules is checked.

Persian homographs

Bijankhan & Moradzade (2004) believe that homographs in Persian appear due to the lack of a one-to-one relationship between phonological and morphological elements and their orthographic correspondence which itself is not rule–based. They classify Persian homographs into two broad categories: lexical (the kind of homographs which are inserted in a dictionary as separate entries) and syntactic (depending on the syntactic context and different derivational and inflectional morphemes which appear in the syntactic context, different homographs are made). In both categories, homographs could be homophones or non–homophones. Then, Bijankhan & Moradzade (2004) classify homographs based on their origin as follows:

Homographs which emerge due to the absence of some diacritics in the Persian writing system. Consider the homograph <فردا>: regardless of the context, it could be pronounced as /fardɒ / (tomorrow) and /fardan/ (individually). This is made due to the absence of the diacritic "Tanwin" (Nunation). If the Tanwin is used, only one of the pronunciations is considered as the correct one: <فرداً> /fardan/ (individually). Homographs which emerge due to the lack of a one-to-one correspondence between graphemes and phonemes in Persian. For example, the grapheme <و> may be pronounced as /v/, /o/ or /u/. So, the word <رود> can have two pronunciations: /rud/ (river) and /ravad/ (go). Homographs which emerge due to the identity of the orthographic and phonological representation of some Persian morphemes including the following:

The morpheme which makes a noun indefinite, a morpheme indicating a noun (place, job, possession, abstractness, diminution, etc.), the inflectional morpheme indicating second person singular in verbs and the derivational morpheme indicating adjectives (subject, object, relationships) all have the same orthographic representation <ی-> /i/ . For example the word <کشاورزی> /keʃ ɒ varzi/, regardless of the context, would mean farming, you are a farmer and a farmer. With regard to this classification, it is worth mentioning that some examples in Homayoonfarrokh's (1985) classification could be classified under the title of homographs. Having studied old Persian, he classifies affix <ی> /j/ into 11 categories including: 1) indicating infinitive structure, so called "esm e ma'xuz", e.g. <سوختگی> /suxtegi/ (the state of being burnt); 2) indicating second person singular in verbs, e.g. <رفتی> /rafti/ (you went); 3) indicating conditional state in verbs accompanying <اگر> /ɂ agar/ (if), e.g. <اگر رفتمی> /ɂ gar raftami/ (if I went); 4) indicating wish, accompanying <کاش> /kɒʃ / (I wish), e.g. <کاش آمدی> /kɒʃ ɂ ɒ madi/ (I wish you came); 5) indicating doubt accompanying <گویا/ گویی> /gujɒ / guji/ (as if), e.g. <گویی پرگوهر دریاستی> /guji por gohar darjɒ sti/ (it's as if you are the sea full of pearls); 6) indicating something happened in dream, e.g. <دیدم به خواب دوش که ماهی برآمدی>

/didam be xɒ b duʃ ke mɒ hi bar ʔ ɒ madi/ (last night, I dreamed the moon rises); 7) making adjectives out of nouns, e.g. <شهر> /ʃ ahr/ (city)+ <ی> /i/ = <شهری> (ʃ ahri/ (urban); 8) indicating continuity in verbs, e.g. <همی گفتی> /hami gofti/ (he was saying); 9) making nouns out of adjectives, e.g. <بزرگ> /bozorg/ (large) + <ی> /i/ = <بزرگی> /bozorgi/ (largness); 10) indicating indefiniteness, e.g. <فردایی> /fardɒ ji/ (one day in the future); 11) indicating worth, e.g. <دیدنی> /didani/ (worth looking). Although some of the mentioned examples in his classification could be considered as the examples of homographs (for example, <بزرگی> regardless of context could mean "you are great", "greatness" and "a great person" or <شهری> could mean "a city" or "urban"), he has not classified them under the title of homographs.The third person singular bound pronoun and one of the morphemes indicating noun have the same orthographic representation <ش-> /eʃ / or /aʃ /. For example, the orthographic form < رویش> may be pronounced as /rujeʃ / (growth) and /rujaʃ / (his/ her face). Sadeghi (1991a,b,c; 1992a,b,c,d,e; 1993a,b,c,d) and Keshani (1992) have also studied the morphological structure of words focusing on Persian derivational morphemes used to form nouns, adjectives and adverbs, but have not referred to homographs. In some homographs, the place of stress distinguishes one form from the other. For example, the orthographic form <ولی> could be pronounced as /va`li/ (but) and /vali`/ (guardian).

Method

A rule-based approach for studying homographs

Word Sense Disambiguation (WSD) is the task of determining which sense of an ambiguous word (word with multiple meanings) is chosen in a particular use of that word, by considering its context (Abed et al: 2015). Up to the present, diverse WSD methods have been proposed. These methods as introduced in Wilks & Stevenson (1998), Montoyo, Suarez, Rigau & Palomar (2005), Bakx (2006), Makki & Homayounpour (2008), Riahi & Sedghi (2012), Singh & Gupta (2015), Mahmoodvand & Hoourali (2015) are overviewed as machine learning (includes supervised and unsupervised) and external knowledge sources. Generally speaking, these methods have the potential limitations. However, almost all methods, without exception, depend on the context in which the ambiguous word occurs (Wang et al: 2013). Word sense ambiguity is also recognized as having a detrimental effect on the precision of information retrieval systems in general and web search systems in particular, due to the sparse nature of the queries involved. Despite continued research into the application of automated word sense disambiguation, the question remains as to whether automated word sense disambiguation with an accuracy below 90% can lead to improvements in retrieval effectiveness; for example, Stokoe, Oakes & Tait (2003) explore the development and subsequent evaluation of a statistical WSD system which demonstrates increased precision from a sense based vector space retrieval model over traditional TF*IDF techniques. Regarding the information retrieval application of WSD, Liu, Yu & Meng. (2005) present a new approach to determine the senses of words in queries using WordNet. In their approach, noun phrases in a query are determined first. For each word in the query, information associated with it, including its synonyms, hyponyms, hypernyms, definitions of its synonyms and hyponyms, and its domains, are used for WSD. By comparing these pieces of information associated with the words in a phrase, it may be possible to assign senses to these words. If the above disambiguation fails, then other query words, if any, are used by going through exactly the same process. If the sense of a query word cannot be determined in this manner,

then a guess is made about the sense of the word in case the guess has at least 50% chance of being correct. If no sense of the word has a 50% or higher chance of being used, then a Web search is applied in the word sense disambiguation process. They claim that based on experimental results, their approach has 100% applicability and 90% accuracy on the most recent robust track of TREC collection of 250 queries. They combine this disambiguation algorithm with their retrieval system to examine the effect of WSD in text retrieval. Experimental results show that the disambiguation algorithm together with other components of the retrieval system yield a result which is 13.7% above that produced by the same system but without the disambiguation, and 9.2% above that produced using Lesk's algorithm. They claim that their retrieval effectiveness is 7% better than the best reported result in the literature. Zhong and Tou Ng (2012) also report successful application of WSD to IR. They have proposed a method for annotating senses to terms in short queries, and also described an approach to integrate senses into an LM approach for IR. In the experiment on four query sets of TREC collection, they have compared the performance of a supervised WSD method and two WSD baseline methods. The experimental results showed that the incorporation of senses improved a state-of-the-art baseline, a stem-based LM approach with PRF method. The performance of applying the supervised WSD method is better than the other two WSD baseline methods. They also proposed a method to further integrate the synonym relations to the LM approaches. With the integration of synonym relations, their best performance setting with the supervised WSD achieved an improvement of 4.39% over the baseline method, and it outperformed the best participating systems on three out of four query sets. Lexical ambiguity is a pervasive problem in natural language processing. However, little quantitative information is available about the extent of the problem or about the impact that is has on information retrieval systems. Krovetz and Croft (1992) report an analysis of lexical ambiguity in information retrieval test collections and on experiments to determine the utility of word meaning for separating relevant documents from non-relevant documents. The experiment show that there is considerable ambiguity even in a specialized database. Word senses provide a significant separation between relevant and non-relevant documents, but several factors contribute to determining whether disambiguation will make an improvement in performance. For example, resolving lexical ambiguity was found to have little impact on retrieval effectiveness for documents that have many words in common with the query.

There exist some WSD studies on Persian homographs which are machine learning-based rather than linguistics-based. For example, Jani and Pilevar (2012) seek to elaborate disambiguation of Persian words with the same written form but different senses using a combination of supervised and unsupervised method which is conducted by means of thesaurus and corpus. Their method is based on a previously proposed one with several differences. These differences include the use of texts which have been collected through supervised or unsupervised methods. In addition, the words of the input corpus were stemmed. In the case of words having different senses and different roles in the sentence, the role of the word in the input sentence was considered for disambiguation. Applying this method to the selected ambiguous words from "Hamshahri", which is a standard Persian corpus, they achieved a satisfactory accuracy of 97 percent in the result. Makki and Homayounpour (2008) describe the disambiguation of Persian homographs in unrestricted texts using thesauri and corpora. The proposed method is based on Yarowsky with some differences. These differences consist of first using collocational information to avoid the

collection of spurious contexts caused by polysemous words in thesaurus categories, and second contribution of all words in the test data context, even those not appeared in the collected contexts to the calculation of the conceptual classes' score. Using a Persian corpus and a Persian thesaurus, this method correctly disambiguated 91.46% of the instances of 15 Persian homographs. This method was compared to three supervised corpus-based methods including Naïve Bayes, Exemplar-based, and Decision List. Unlike supervised methods, this method needs no training data, and has a good performance on the disambiguation of uncommon words. In addition, this method can be used to remove some kinds of morphological ambiguities. Riahi and Sedighi (2012) believe that supervised methods are the most common solutions for WSD. However, they need large tagged corpora which are not available in some languages such as Persian. The Semi-Supervised methods can solve this problem by using a small tagged corpus and a large untagged corpus. Riahi and Sedighi (2012) present a coarse-grained work in WSD that uses tri-training as the semi-supervised method and decision list as supervised classifier for training. The proposed method was evaluated on a corpus and was reported as more precise than the conventional decision list when the tagged corpus is small.

The present study is a corpus-based approach to WSD which benefits from POS tagging. A corpus-based approach extracts information regarding the frequency of homographs from a large annotated data collection, referred to as a POS-tagged corpus. The possible means of attributing the right POS tag to ambiguous words is to extract the homographs with high frequency in the corpora, then introducing a method based on the distributional information and context to disambiguate the POS tag of the mentioned homographs. Unlike the previous studies, the proposed method in this paper is a combination of machine learning approach to search for homographs in corpora as well as checking the accuracy of extracted rules, and the linguistic approach for studying homograph in linguistic contexts to extract context-sensitive rules for allocating the right POS tag to the studied homographs. Since we needed tagged corpora to search for homographs, we had to use the three available corpora including the Persian written corpus: Peykare, known as Bijankhan corpus (Bijankhan et al. 2011), the Farsi linguistic database, known as paygah-e dadegan-e zaban-e Farsi (Assi: 1997) and the Persian syntactic dependency Treebank (Rasooli et al: 2013). Search tools are used to look for homographs in two Persian tagged corpora (Peykare and syntactic dependency Treebank). The search tool looks at each word and its tag(s) in the corpus and finds words with more than one POS tag. For example, the search tool of "Peykare/ Bijankhan" corpus, operates in the following way:

Each row in Peykare includes one word and its POS. There is a set named "dictionary" structured such that that every word together with its POS(s) is saved in the set. The search tool studies each row of the corpus; if the word in the row is absent in the dictionary, it adds the word and its POS tag to the dictionary. If the word already exists in dictionary, the program studies whether the inserted POS of the word in the row has already been inserted for this word in dictionary or not. If not, it adds the new tag to the list of POS tags of this word to the dictionary. Finally, the search tool studies the whole dictionary and the words with more than one POS tag are listed as the output.

A general study of the list of homographs shows that the number of homographs in different Persian corpora is considerable which means that POS tag disambiguation is necessary, otherwise text processing would face problems. The study shows that most of these

homographs emerge as a result of the same orthographic representation for some inflectional and derivational morphemes including the morpheme indicating the indefiniteness of the noun, the noun maker morpheme (indicating place, job, possession, diminution and abstractness), the second person singular morpheme in verbs and the adjective maker morpheme (indicating the subject, object and relation) all having the orthographic representation <ی> /i/. So, the result shows that the most frequent homographs in corpora are noun and adjective homographs ending in <ی> /i/. Such homographs could be considered as the main source of ambiguity in the texts. Only the context can distinguish the tag of such homographs. For example, the word <کشاورزی> /keshavarzi/, a kind of homograph ending in <ی>, would mean farming or a farmer regardless of the context. The POS tag disambiguation of such homographs can make word sense disambiguation easier and lead to better text processing.

After extracting the most frequent homographs in the corpora (noun and adjective homographs ending in <ی> /i/), a list of such homographs in the syntactic context was extracted (using the knowledge of neighboring words). Unlike the previous studies, including that of Homayoonfarrokh (1985), in which the related contexts were not considered for studying suffixes like <ی> (because the aim of the study was not word sense/tag disambiguation), the present study considers the context as the main factor for word tag disambiguation. The context composed of the words found to the right and/or the left of a certain word, thus collocational or co-occurrence information was considered. In the present study, homographs ending in <ی-> were studied with regard to a history of 10 windows (considering 10 left context words and tokens (including delimiters) and 10 right context words and tokens (including delimiters)) in order to decide on the right POS tag of the homograph based on the structure of the sentence. A rule-based program was used to make a list of noun and adjective homographs ending in <ی> which runs using Python. This program uses a tagged corpus, in this case, the Bijankhan corpus, and searches for any tagged word which ends with <ی>, then the word with its context (10 words before and after the studied word) is presented. For example, considering the homograph <درمانی> one of the context in which this homograph is used is as follows:

خرده‌فرهنگی، ما روی شیوهٔ تفسیر علایم بیماری و نیز درمانی که به دنبال آن هستیم، تاثیر می‌گذارند. با

The related POS tag of each word in this context is also presented:

خرده‌فرهنگی (ADJ) / adjective	؛ (DELM) / punctuation	ما (PRO)/ pronoun
روی (P) / preposition	شیوهٔ (N) /noun	تفسیر (N) /noun	علایم (N) /noun
بیماری (N) /noun	و (CON) / conjunction	نیز(CON) / conjunction	درمانی (N) /noun
که (CON) / conjunction	به (P)/ preposition	دنبال (N) / noun	آن (PRO)/ pronoun
هستیم (V)/ verb	، (DELM) / punctuation	تاثیر (N) / noun	می‌گذارند (V)/ verb
. (DELM) / punctuation	با (P)/ preposition		

So, 10 orthographic forms (including words and punctuation marks) before each homograph and 10 orthographic forms after each homograph are presented, all of which are accompanied by the related POS tags.

One such study is presented in Table 1 (the actual file is an Excel sheet, so only 3 or 4 words before and after the homograph is presented here because of a lack of space).

Table 1
An example of homographs ending in <ی> in the syntactic context in which the neighboring words are tagged

نامشخص	اغلب	ويروسي(viral)	بيماريهاي	بالاخص	عفوني
ADJ_SIM	ADV_NI_Q_SIM	ADJ_SIM	N_PL_COM_GEN	CON_GMC	ADJ_SIM
و	نيستند	ويروسي	لزوماً	كه	مي دهد
CON_GMC	V_PRS_NEG_6	ADJ_SIM	ADV_NI_NQ_SIM	CON_RELC	V_PRS_POS_3
.	پرداختند	ويروسي	بيماريهاي	درمان	در
DELM	V_PA_SIM_POS_6	ADJ_SIM	N_PL_COM_GEN	N_SING_COM_GEN	P_GENR
.	مي گشايند	ويروسي	بيماريهاي	درمان	براي
DELM	V_PRS_POS_6	ADJ_SIM	N_PL_COM_GEN	N_SING_COM_GEN	P_GENR_GEN
كرد	استفاده	ويروسي	بيماريهاي	عليه	،
V_PA_SIM_POS_3	N_SING_COM	ADJ_SIM	N_PL_COM_GEN	P_GENR_GEN	DELM
كه	است	ويروسي(a virus)	منشأ	به	نزديك
CON_RELC	V_PRE_SIM	N_SING_COM_INYA	N_SING_COM_GEN	P_GENR	N_SING_COM
كه	را	ويروسي	،	آمريكايي	پژوهشگران
CON_RELC	P_DEFI	N_SING_COM_INYA	DELM	ADJ_SIM	N_PL_COM_GEN
كه	است	ويروسي	نخستين	اين	.
CON_RELC	V_PRE_SIM	N_SING_COM_INYA	ADJ_SUP	DET	DELM
شما	كه	ويروسي	مراقب	،	مي كنيد
PRO_DEF_NR_NIP_2	CON_RELC	N_SING_COM_INYA	N_SING_COM_GEN	DELM	V_PRS_POS_5
به	موسوم	ويروسي	آنتي ژن	سطحي	شاخصهاي
P_GENR	ADJ_SIM	N_SING_COM_INYA	N_SING_COM_GEN	ADJ_SIM_GEN	N_PL_COM_GEN

So, we have the words ending in <ی> with a history of 10 surrounding words. It means that 10 words before the homograph ending in <ی> and 10 words after it are presented. Below every word (as in table 1) there is the related POS tag of the word. For example, under the

word <که>, its related POS tag (CON, which means conjunction) is presented. Then the result, which consisted of millions of words in contexts, was studied to extract context-sensitive rules for allocating the right POS tag to the homograph in syntactic structures. The extracted rules include the following ones: (note: unlike English, the Persian writing system is from right to left)

1. a. Preposition (P) + (Quantifier (QUA)) + Noun (N)

This rule means that if a word ending in <ی> is preceded by a preposition and an optional quantifier, the POS tag of that word is NOUN.

1. b. Preposition (P) + (Quantifier (QUA)) + Noun (N) + Conjunction (CON) + Noun (N)

This rule means that if a word ending in <ی> is preceded by a preposition and an optional quantifier, then a noun and a conjunction, the POS tag of that word is NOUN.

Such rules were checked in lots of words in the mentioned context. It means that to check whether such a rule is verified or not, lots of nouns ending in <ی> were studied in similar contexts. Examples include:

(بهP) هیچ (QUA)ارتشی (N....نمی‌دهد اجازه نیز)

To (P) any (QUA) army (N) as well (ADV) let /V/

'He does not let any army as well

(بهP) دوست (N) و (CON) خویشاوندی (N.....بود گفته)

To (p) friend (N) and (CON) a relative (N) had told

He had said to a friend and a relative ….

2. a. Preposition (P) + words meaning “kind/type/form” + Adjective (ADJ)

This rule means that if a word ending in <ی> is preceded by a preposition and words meaning “kind/type/form”, then the POS tag of that word is ADJECTIVE.

2. b. Preposition (P) + words meaning “kind/type/form” (surat/lahaz/nazar/noʔ) + Adjective (ADJ)+ Conjunction (CON) + Adjective (ADJ)

This rule means that if a word ending in <ی> is preceded by a preposition and words meaning “kind/type/form”, then a Noun and a conjunction, the POS tag of that word is ADJECTIVE.

Such rules were checked in lots of words in the mentioned context. It means that to check whether such a rule is verified or not, lots of nouns ending in <ی> were studied in a similar context. Examples include:

(بهP) صورت رقمی (ADJاندشده ضبط)

In (P), the form of numerical (ADJ) have been recorded.

“They have been recorded in the numerical form.”

(بهP) صورت تستی (ADJ) یا (CON) نوشتاری (ADJ.....)

In the form of multiple choices (ADJ) or (CON) written (ADJ)

3. Word meaning “as” (be ʔ onvan e) + (superlative adjective (adj-SUP)) + Noun (N)

This rule means that if a word ending in <ی> is preceded by a word meaning “as” (be ʔ onvan e) and an optional superlative adjective, the POS tag of that word is NOUN.

Such a rule was checked in lots of words in the mentioned context. It means that to check whether such a rule is verified or not, lots of nouns ending in <ی> were studied in a similar context. Examples include:

(به عنوان جوان‌ترینADJ-SUP) کارگردانی (N.....که)
As the youngest (ADJ-SUP) director who….

4. Preposition (P) + Noun (N) + preposition (p)

This rule means that if a word ending in <ی> is preceded by a preposition and followed by another preposition, the POS tag of that word is NOUN.

Such a rule was checked in lots of words in the mentioned context. It means that to check whether such a rule is verified or not, lots of nouns ending in <ی> were studied in similar context. Examples include:

(برایP) تاثیرگذاری (N) بر (P.....اعضای)
For (P), impacting (N) on (P) the members of ….

5. a. A word indicating time periods such as: dore/asr/doran/zaman/sal/senin + Noun (N)

This rule means that if a word ending in <ی> is preceded by a word indicating time periods such as: dore/asr/doran/zaman/sal/senin, the POS tag of that word is NOUN.

5. b. preposition (P) + a word indicating time periods such as: dore/asr/doran/zaman/sal/senin + Noun (N)

This rule means that if a word ending in <ی> is preceded by a preposition, then a word indicating time periods such as: dore/asr/doran/zaman/sal/senin, the POS tag of that word is NOUN.

Such rules were checked in lots of words in the mentioned context. It means that to check whether such a rule is verified or not, many nouns ending in <ی> were studied in a similar context. Examples include:

(نسبت می‌دهندNبه مشکلات زمان کودکی)
To problems time period of childhood (N) relate.
“They relate it to childhood problems”
(ازP) دوران نوجوانی (N.....تا هنگام وفات)
From (P) the time of teenage (N) till death
“since his teenage years till his death”

6. a. A quantifier meaning any/every + (words meaning kind of (noʔ /gune)) + (number) + Noun (N)

This rule means that if a word ending in <ی> is preceded by a quantifier meaning any/every, then an optional word meaning kind of (noʔ / gune) or number, the POS tag of that word is NOUN.

6. b. A quantifier meaning any/every + (words meaning kind of (noʔ / gune)) + (number) + Noun (N) + conjunction (CON) + Noun (N)

This rule means that if a word ending in <ی> is preceded by a quantifier meaning any/every, then an optional word meaning kind of (noʔ / gune) or number, then a noun and conjunction, the POS tag of that word is NOUN.

Such rules were checked with lots of words in the mentioned context. It means that to check whether such rule is verified or not, many nouns ending in <ی> were studied in similar contexts. Examples include:

(هرQUA) نوع آب و هوایی (N.....)

Any kind of whether (N)

(از زندگی‌مان....N) نکته‌ای (CON) یا (N) بخشی (QUAهیچ)

No (QUA) part (N) or (CON) point (N) of our life

7. verb (V) + conjunction (CON) + Noun (N) + preposition (P)

This rule means that if a word ending in <ی> is preceded by a verb, then a conjunction and followed by a preposition, the POS tag of that word is NOUN.

Such a rule was checked in lots of words in the mentioned context. It means that to check whether such a rule is verified or not, lots of nouns ending in <ی> were studied in similar contexts. Example include:

(او به وجود می‌آیدP) در (N) احساساتی (CON) و (Vتماشا می‌کند)

Looks (V) and (CON) some feeling (N) in (P) him arises.

"…looks and some feelings arise in him"

8. Verb (V) + conjunction (CON)/Punctuation (,) (DELM) + Noun (N) + (verb (V)) + conjunction meaning "that"

This rule means that if a word ending in <ی> is preceded by a verb, then a conjunction or punctuation (,) Punctuation (DELM) and is followed by an optional verb and conjunction meaning "that," the POS tag of that word is NOUN.

Such a rule was checked in lots of words in the mentioned context. It means that to check whether such a rule is verified or not, lots of nouns ending in <ی> were studied in a similar context. Examples include:

(انجام می‌داد V) ، (DELM) دوره‌ای (N....که تهران اقتصاد روستایی داشت)

Was doing (V), (DELM) period (N) that Tehran had rural economy

9. Adjective (ADJ) + conjunction (CON) + Adjective (ADJ)

This rule means that both sides of a conjunction should be the same, two adjectives can be inserted: one before the conjunction and the other after the conjunction.

Such a rule was checked in lots of words in the mentioned context. It means that to check whether such a rule is verified or not, many nouns ending in <ی> were studied in a similar context. Examples include:

(موضع طبیعیADJ) یا (CON) قراردادی (ADJ.... خود خارج)

Condition natural (ADJ) or (CON) conventional (ADJ) his out…

"His natural or conventional condition …"

10. Noun (N) + Adjective (ADJ) + conjunction (CON) + Noun (N) +Adjective (ADJ)

This rule means that both sides of a conjunction should be the same, two noun clauses (noun + adjective) can be inserted: one before the conjunction and the other after the conjunction.

Such a rule was checked in lots of words in the mentioned context. It means that to check whether such a rule is verified or not, many nouns ending in <ی> were studied in a similar context. Examples include:

(همان تغییرN) اجتماعی (ADJ) و (CON) تغییر (N) فرهنگی (ADJ..... لازم)

The same change (N) social (ADJ) and (CON) change (N) cultural (ADJ) necessary….

"The same necessary social and cultural change …."

11. Noun (N) + adjective (ADJ) + Adjective (ADJ)

This rule means that if a word ending in <ی> is preceded by an adjective and the adjective is preceded by a noun, the POS tag of that word is ADJECTIVE. Examples include:

(ریاضیدانN) بزرگ (ADJ) نیشابوری (ADJ)

Mathematician (N) Ezafe great (ADJ) Ezafe from Neyshaboor (ADJ)

'The great mathematician from Neyshaboor '

In the above example, Ezafe means: The elements within a noun phrase or adjective phrase are linked by the enclitic particle called Ezafe. This morpheme is usually an unwritten vowel, but it could also have an orthographic realization in certain phonological environments. In most cases, this relation can be translated as a genitive structure. Examples of this construction are given below (Megerdoomian 2000):

a. sedâ-ye pâ-ye man

sound-ez foot-ez my

‘(the) sound of my footsteps’

b. ru-ye miz

on-ez table

‘on the table’

12. Noun (N) + adverb (ADV) + Adjective (ADJ)

This rule means that if a word ending in <ی> is preceded by an adverb and the adverb is preceded by a noun, the POS tag of that word is ADJECTIVE. Examples include:

(هوایN) بسیار (ADV) سردی (ADJ است)

weather (N) Ezafe very (ADV) cold (ADJ) is

'It’s a very cold whether '

13. Demonstrative adjective (<این> /ɔ in/, <آن> /ɔ ɒ n/) + (<نوع> /noɔ / a word meaning kind of) + Noun (N)

This rule means that if a word ending in <ی> is preceded by an optional word, meaning kind of which itself is preceded by a demonstrative adjective in Persian, then the POS tag of that word is NOUN. Examples include:

(این(demonstrative adjective) نوع زندگی N)

ɔ in/ (demonstrative adjective) /noɔ / /zendegi/ (N)

this (demonstrative adjective) kind living (N)

“This kind of living’

Thirty-six context-sensitive rules were extracted from the corpus. Then, the accuracy of the rules was checked via programming. The result is presented in Table 2.

Table 2

The result of checking the accuracy of 36 context-sensitive rules

A: Rule Name	B: All_Count	C: True_Count	D: True percent	E: False_Count	F: False_Percent
rule01a	8849	8320	94.0219234	529	5.97807662
rule01b	1776	1493	84.0653153	283	15.9346847
rule02a	692	421	60.8381503	271	39.1618497
rule02b	63	54	85.7142857	9	14.2857143
rule03	222	212	95.4954955	10	4.5045045
rule04	1767	1625	91.9637804	142	8.03621958
rule05a	499	296	59.3186373	203	40.6813627
rule05b	310	193	62.2580645	117	37.7419355
rule06a	643	626	97.3561431	17	2.64385692

A: Rule Name	B: All_Count	C: True_Count	D: True percent	E: False_Count	F: False_Percent
rule06b	109	101	92.6605505	8	7.33944954
rule07	1085	597	55.0230415	488	44.9769585
rule08	546	470	86.0805861	76	13.9194139
rule09a	4950	4071	82.2424242	879	17.7575758
rule09b	6305	4180	66.29659	2125	33.70341
rule10a	2138	1561	73.0121609	577	26.9878391
rule10b	1983	1428	72.0121029	555	27.9878971
rule11	1320	1147	86.8939394	173	13.1060606
rule12a	51	5	9.80392157	46	90.1960784
rule12b	34	30	88.2352941	4	11.7647059
rule13	407	348	85.5036855	59	14.4963145
rule14	12550	4930	39.2828685	7620	60.7171315
rule15a	6631	5562	83.8787513	1069	16.1212487
rule15b	1154	953	82.5823224	201	17.4176776
rule16	35	32	91.4285714	3	8.57142857
rule17	842	655	77.7909739	187	22.2090261
rule18	3209	1670	52.0411343	1539	47.9588657
rule19	38	24	63.1578947	14	36.8421053
rule20	1205	621	51.5352697	584	48.4647303
rule21a	1378	1100	79.8258345	278	20.1741655
rule21b	2280	865	37.9385965	1415	62.0614035
rule22a	8818	4936	55.9764119	3882	44.0235881
rule22b	5066	4090	80.7343071	976	19.2656929
rule23a	1371	1205	87.8920496	166	12.1079504
rule23b	169	162	95.8579882	7	4.14201183
rule24	718	608	84.6796657	110	15.3203343
rule25	143	139	97.2027972	4	2.7972028
rule26	387	327	84.496124	60	15.503876
rule27	243	239	98.3539095	4	1.64609053
rule28	3522	2125	60.3350369	1397	39.6649631
rule29a	425	246	57.8823529	179	42.1176471
rule29b	91	37	40.6593407	54	59.3406593
rule30	729	555	76.1316872	174	23.8683128
rule31	530	433	81.6981132	97	18.3018868
rule32	361	242	67.0360111	119	32.9639889
rule33a	4691	1807	38.5205713	2884	61.4794287
rule33b	130	92	70.7692308	38	29.2307692
rule34a	2308	1633	70.7538995	675	29.2461005
rule34b	399	283	70.9273183	116	29.0726817
rule35	78	77	98.7179487	1	1.28205128
rule36	167	118	70.6586826	49	29.3413174

(Explanation: Rule 1.a. shows that the number of homographs about which this rule is worth studying is 8849, the rule is applicable in 8320 cases (the number of the true-count) and amounts to %94.02 (the true percentage), and in 529 cases the rule is not applicable (the false-count) which amunts to %5.97 (the false percentage).

The result showed that the accuracy of most rules is high which proves most rules are true.

Discussion

Since homographs are one of the main challenges faced in text processing, the frequency of homographs was studied in a number of Persian corpora to extract the most frequent homographs. Search tools were used to search for homographs in the Persian corpora and a lengthy list of homographs was extracted. Making a list of homographs has two main functions: 1. the list indicates that the number of homographs in the Persian corpora is high which means that word POS tag disambiguation is necessary, otherwise text processing would face problems. 2. The homographs with high frequency (homographs made as a result of the same orthographic representation of some inflectional and derivational morphemes including: the inflectional morpheme indicating the indefiniteness of the noun, the noun maker morpheme, the second person singular morpheme in verbs and the adjective maker morpheme) can be used for word POS tag disambiguation using the syntactic context to specify the correct POS tag for them in the corpus. Based on the list, the most frequent homographs are nouns and adjectives ending in <ی>. The POS tag disambiguation of such homographs can make word sense disambiguation easier and lead to better text processing. In this part of the study, a list of noun and adjective homographs ending in <ی> in syntactic contexts is made (using knowledge of neighboring words in which homographs ending in <-ی> were studied with regard to a history of 10 windows (before and after each homograph) in order to decide about the right POS tag of the homograph based on the structure of the sentence. Then, the result was studied to extract context-sensitive rules for allocating the right POS tag to the homograph in syntactic structures. Afterwards, the accuracy of rules was checked via programming. The result showed that the accuracy of most rules is high which proves most rules are true.

References

Abed. SA. Tiun, S. and Omar, N. 2015. Harmony Search Algorithm for Word Sense Disambiguation. PLoS ONE 10(9): e0136614.

Alayiaboozar, E. & Bijankhan, M. (2013). Persian orthographic depth. *Journal of language researches*. 4(1), 1-19. [in Persian]

Assi, S. M. (1997). Farsi linguistic database (FLDB). *International journal of Lexicography*, 10(3), *EURALEX Newsletter*, p.5.

Bakx, G. E., (2006). Machine learning techniques for Word Sense Disambiguation. Ph D theses, Universitat Politecnica de Catalunya. Barcelona.

Bijankhan, M. & Moradzadeh, S.h.. 2004. Homographs in Persian morphology. *Proceeding of the first workshop on the Persian language and computer*. Tehran University .53–63. [in Persian]

Bijankhan, M., Sheykhzadegan, J., Bahrani, M. and Ghayoomi, M. (2011). Lessons from Building a Persian Written Corpus: Peykare. *Language Resources and Evaluation*, 45(2), 143–164.

Gaustad, T. 2004. Linguistic Knowledge and Word Sense Disambiguation. Netherlands Organization for Scientific Research. Phd. Dissertation.

Gottlob, L.R., Goldinger, S.D., Ston, G.O., & Orden, G.C.V. 1999. Reading homographs: orthographic, phonologic and semantic dynamics. *Journal of Experimental Psychology: Human Perception and Performance*. 25(2), 561-574.

Homayoonfarrokh, A. (1985). Comprehensive Grammar of Persian. Aliakbar Elami Publication. Scientific Press Institution. [in Persian]

Jani.F. & Pilevar, A.H. (2012). Word Sense Disambiguation of Persian Homographs. *In Proceedings of the 7th International Conference on Software Paradigm Trends (ICSOFT-2012).* 328-331.

Keshani , Kh. 1992. Derivational suffixes in modern Persian. Markaz Nashr e Daneshgahi. Tehran. [in Persian]

Krovetz, R. & Croft, W. B. (1992). Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems (TOIS)*, 10 (2). 115-141.

Liu, S.H., Yu, C. & Meng, W. 2005. Word Sense Disambiguation in Queries. *In Proceedings of the 14th ACM international conference on Information and knowledge management*. 525-532

Mahmoodvand, M & Hourali, M. 2015. Persian Word Sense Disambiguation Corpus Extraction Based on Web Crawler Method. *ACSIJ Advances in Computer Science: an International Journal*, 4(5), 101-106.

Makki R., & Homayounpour M. M. (2008). Word Sense Disambiguation of Farsi Homographs Using Thesaurus and Corpus. Nordström B., Ranta A. (Eds) *Advances in Natural Language Processing. Lecture Notes in Computer Science*, Vol. 5221. Springer, Berlin, Heidelberg.

Megerdoomian, K. (2000). Unification - based Persian morphology. *In Proceedings of CICLing*. Centro de investigacion en computacion - IPN, Mexico. 311 – 318.

Merriam Webster.com.2019. "homographs".http://www.merriam-webster.com (3 July 2019)

Montoyo, A., Suarez, A., Rigau, G., and Palomar, M. (2005). Combining knowledge –and corpus –based Word-Sense-Disambiguation methods. *Journal of Artificial Intelligence Research*. 23: 299-330.

Rasooli, M. S., Kouhestani, M., and Moloodi, A. (2013). Development of a Persiansyntactic dependency treebank. InProceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 306–314, Atlanta, Georgia. Association for Computational Linguistics.

Riahi, N., & Sedghi, F. (2012). A semi- supervised method for Persian homograph disambiguation. *Proceedings of 20th Iranian conference on Electrical Engineering. ICEE*. University of Tehran. Iran. [in Persian]

Sadeghi, A. (1991a). Ways and possibilities of word formation in modern Persian. 1. Nashr e Danesh. 64. 12-18. [in Persian]

Sadeghi, A. (1991b). Ways and possibilities of word formation in modern Persian. 2. Nashr e Danesh. 65. 6-12. [in Persian]

Sadeghi, A. (1991c). Ways and possibilities of word formation in modern Persian. 3. Nashr e Danesh. 67. 28-33. [in Persian]

Sadeghi, A. (1992a). Ways and possibilities of word formation in modern Persian. 4. Nashr e Danesh. 69. 21-25. [in Persian]

Sadeghi, A. (1992b). Ways and possibilities of word formation in modern Persian. 5. Nashr e Danesh. 70. 39-45. [in Persian]

Sadeghi, A. (1992c). Ways and possibilities of word formation in modern Persian. 6. Nashr e Danesh. 71. 15-19. [in Persian]

Sadeghi, A. [1992d]. Ways and possibilities of word formation in modern Persian. 7. Nashr e Danesh. 72. 19-23. [in Persian]

Sadeghi, A. (1992e). Ways and possibilities of word formation in modern Persian. 8. Nashr e Danesh. 74. 22-29. [in Persian]

Sadeghi, A. (1993a). Ways and possibilities of word formation in modern Persian. 9. Nashr e Danesh. 75. 9-15. [in Persian]

Sadeghi, A. (1993b). Ways and possibilities of word formation in modern Persian.10. Nashr e Danesh. 76. 15-23. [in Persian]

Sadeghi, A. (1993c). Ways and possibilities of word formation in modern Persian. 11. Nashr e Danesh. 77. 21-25. [in Persian]

Sadeghi, A. (1993d). Ways and possibilities of word formation in modern Persian. 12. Nashr e Danesh. 79, 80. 12-15. [in Persian]

Singh, H., & Gupta, V. 2015. An insight to word sense disambiguation techniques. *International journal of computer applications*, 118 (23), 32-39.

Stokoe, Ch., Oakes, M., and Tait, J. 2003. Word sense disambiguation in information retrieval revisited. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*. 159-166.

Wang. X., Zuo, W., and Wang, Y. 2013. A Novel Approach to Word Sense Disambiguation Based on Topical and Semantic Association. The Scientific World Journal. Hindawi Publishing Co, UK.

Wilks, Y & Stevenson, M. (1998). Word sense disambiguation using optimized combinations of knowledge sources. *Proceedings of the 17th international conference on computational linguistics and the 36th annual meeting of the association for computational linguistics (COLING-ACL` 98).* Montreal, Canada. Pp 1

Zeroual, I., Lakhouaja, A. & Belahbib, R. (2017). Towards a standard part of speech tagset for the Arabic language. *Journal of King Saud University –Computer and Information Sciences*, 29(2), 171-178.

Zhong, Zh, & Tou Ng, H. (2012). Word sense disambiguation improves information retrieval. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*: Volume 1. 273-282.