A Probabilistic Approach to Persian Ezafe Recognition

Habibollah Asghari
Department of ECE,
University of Tehran,
Tehran, Iran
habib.asghari@ut.ac.ir

Jalal Maleki
Department of CIS,
Linköpings Universitet
SE-581 83 Linköping, Sweden
jalal.maleki@liu.se

Heshaam Faili
Department of ECE,
University of Tehran,
Tehran, Iran
hfaili@ut.ac.ir

Abstract

In this paper, we investigate the problem of Ezafe recognition in Persian language. Ezafe is an unstressed vowel that is usually not written, but is intelligently recognized and pronounced by human. Ezafe marker can be placed into noun phrases, adjective phrases and some prepositional phrases linking the head and modifiers. Ezafe recognition in Persian is indeed a homograph disambiguation problem, which is a useful task for some language applications in Persian like TTS. In this paper, Part of Speech tags augmented by Ezafe marker (POSE) have been used to train a probabilistic model for Ezafe recognition. In order to build this model, a ten million word tagged corpus was used for training the system. For building the probabilistic model, three different approaches were used; Maximum Entropy POSE tagger, Conditional Random Fields (CRF) POSE tagger and also a statistical machine translation approach based on parallel corpus. It is shown that comparing to previous works, the use of CRF POSE tagger can achieve outstanding results.

1 Introduction

In Persian language, Ezafe is an unstressed short vowel /-e/ (or /-ye/ after vowels) which is used to link two words in some contexts. Although Ezafe is an important part of the Persian phonology and morphology, it does not have a specific character representation, and so is not usually written. However, it is pronounced as a short vowel /e/. Sometimes, for disambiguation purposes it is preferred to explicitly mark its presence by a written symbol (the diacritic Kasre) after some words in order to facilitate the correct pronunciation.

The most important application of Ezafe recognition is a text to phoneme tool for Text To Speech (TTS) Systems. Other application of Ezafe recognition is identifying the dependency of a word in a Noun Phrase. (Oskouipour, 2011, Mavvaji and Eslami, 2012)

In this research, we would like to investigate various approaches to correctly recognize genitive cases in Persian language. Shortly, the contributions of this paper are as follow:

- Modeling the Ezafe recognition task as a sequence labeling system.
- Using HMM and CRF as sequence labelers.
- Modeling the Ezafe recognition task as a monotone translation problem which can be tackled by phrase based SMT approach.
- Using a big amount of test and gold data, so the results are considerably reliable.
- To enhance the results of the system, five Persian-specific features which discriminate the results in high-precision low-recall fashion, have been proposed.
- The recognition rate has achieved outstanding results in comparison to the previous works.

This task is closely related to the task of determining short vowels in Arabic language. So, although the aim of this paper is to recognize Ezafe in Persian language, but all the methods investigated here is applicable to determine short vowels in Arabic language.

In the next section a clear definition of the problem is presented and the characteristics of Persian language are introduced. In Section 3 we will give a precise definition of Ezafe. Section 4 provides an overview of previous works on Ezafe recognition. Our approach will be described in Section 5 followed by two sections including corpus selection process and implementation of proposed method. Conclusion and recommendations for future works will be presented in the last section.

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 138–142,
Gothenburg, Sweden, April 26-30 2014.

2 An Overview of Persian Language

Persian Language belongs to Arabic script-based languages. This category of languages includes Kurdish, Urdu, Arabic, Pashtu and Persian. They all have common scripting, and somehow similar writing system.

In Arabic script-based languages, the most common features are absence of capitalization, right to left direction, lack of clear word boundaries, complex word structure, encoding issues in computer environment, and a high degree of ambiguity due to non-representation of short vowels in writing (Farghaly, 2004). Note that Ezafe recognition and homograph disambiguation problem mostly deals with the last mentioned feature.

One of the problems in Persian language processing is long-distance dependencies. This phenomenon complicates Ezafe recognition task even for humans (Ghomeshi, 1996). Another problem is how to determine phrase/word boundaries. In Persian language, affixes can be written in three formats; completely separated by a space delimiter, separated by half-space[1], or can be attached to its main word. So, determining word and phrase boundaries are somehow a complicated task in Persian.The third challenge arises by pronoun drop due to the morphology of Persian language.

3 Ezafe Definition

Historically, Persian Ezafe had a demonstrative morpheme in old Iran (Estaji and Jahangiri, 2006). It was related to a demonstrative /hya/, which links the head noun to adjectival modifiers, to the possessor NP (Samvelian, P., 2007). In evolution of Persian language, /hya/ became /–i/ in Middle Persian and progressively lost its demonstrative value to end as a simple linker. In recognizing Ezafe, we should consider all morphological, syntactic, semantic and discourse views (Parsafar, 2010). It should be noted that Ezafe can be iterated within the NP, occurring as many times as there are modifiers.

4 Previous Works

As a first attempt to recognize Ezafe in Persian text, Bijankhan (Bijankhan, 2005) used a pattern matching algorithm for Ezafe recognition. He has used POS tags and also semantic labels (such as place, time, ordinal numbers ...) to obtain a statistical view of Ezafe markers. He manually derived 80 most frequent patterns such as Noun-Noun and Noun-Adjective etc. The most frequent combinations were extracted based on a 10 million-wordscorpus.

In a research accomplished by (Isapour, et al., 2007), the researchers rely on the fact that Ezafe can relate between head and its modifiers so as to help to build NPs. So by parsing sentences and finding Phrase borders, the location of Ezafe in the sentence can be found. In this work, the sentences were analyzed using a Probabilistic Context Free Grammar (PCFG) to derive phrase borders. Then based on the extracted parse tree, the head and modifiers in each phrase can be determined. In the last phase, a rule based approach was also applied to increase the accuracy in Ezafe marker labeling. For training the algorithm, 1000 sentences were selected and a parse tree was built for each of them. Because of the limited number of parsed sentences for training, the results cannot be extended for general applications.

There were also other attempts to effectively recognize Ezafe marker in Persian text, such as (Zahedi, 1998) based on fuzzy sets. Also, (Oskouipour, 2011) developed a system based on Hidden Markove Model to correctly identify Ezafe markers. (Mavvaji and Eslami, 2012) had another attempt by syntactic analysis. There are also some implementations using neural networks (Razi and Eshqi, 2012). Some of the results can be seen in Table 4.

5 Our Approach

In this paper we have investigated two types of POS taggers, and also a MT-based approach. In the following section, these approaches will be explained and the results will be compared to previous work.

A. Ezafe recognition as a POS tagging problem

Part Of Speech tagging is an effective way for automatically assigning grammatical tags to words in a text. In Persian, POS tagging can be applied as a homograph disambiguation problem for correct pronunciation of words in a sentence (Yarowsky, 1996). There are powerful POS tagger algorithms such as statistical, rule based, transformation based and memory based learning methods. In this research we have used two schemes of statistical POS tagging for Ezafe recognition. The first one is a Maximum Entropy tagger that has been investigated by (Toutanova and Manning. 2000) and (Toutanova, et al.

[1]A Non-Joint Zero Width (NJZW) letter

2003). In order to implement this approach, we have used Stanford toolkit as a MaxEnt tagger. The second approach is based on Conditional Random Fields (CRF) model, that was first introduced by (Lafferty, et al., 2001) and then (Sha and Pereira. 2003).

B. Ezafe recognition as a translation problem

We can consider the Ezafe recognition problem as a monotone translation problem. In other words, it can be considered as a noisy channel problem. The original training text without the Ezafe marker can be used as source language, and the tagged text can be used as destination language. So, we can apply these parallel corpora as inputs to a phrase-based Statistical Machine Translation (SMT) system.

In the experiments, we have used monotone SMT with distortion limit equal to zero. For implementing SMT, we have used Moses toolkit. It should be mentioned that in the case of Ezafe recognition, we can use a SMT system without re-ordering. By using phrase-based SMT, the local dependencies between the neighboring words are handled by the phrase table. Also some of the dependencies between different phrases can be tackled by the language model.

6 Data Preparation

In this work, we have used Bijankhan corpus (Bijankhan, 2004, Amiri, et al, 2007). The content of this corpus is gathered from daily news and common texts, covering 4300 different subjects. It contains about 10 million tagged words in about 370000 sentences. The words in the corpus have been tagged by 550 tags based on a hierarchical order, with more fine-grained POS tags like 'noun-plural-subj'. About 23% of words in the corpus are tagged with Ezafe. We have used an extended version of POS tags, named POSE (Part of Speech tags + Ezafe tag) that can be constructed by adding Ezafe markers to original first level tags. Table 1 shows the statistics of POSE tags.

POSE	Frequency	% in Ezafe markers	% in all corpus
N-e	*1817472*	*81.87*	*18.39*
ADJ-e	*223003*	*10.05*	*2.26*
P-e	*111127*	*5.01*	*1.125*
NUM-e	*27860*	*1.26*	*0.28*
others	*40477*	*1.81*	*0.41*
Total	*2219939*	*100 %*	*22.46*

Table 1 - Ezafe Statistics in Bijankhan Corpus

7 Performance Metrics

The ordinary measures that can be used based on confusion matrix are Precision, Recall and F1 measure. Another measure that can be used in this binary classification problem is Mathews Correlation Coefficient (MCC). This measure indicates the quality of the classifier for binary class problems especially when two classes are of very different sizes.We have also considered two other measures; true positive rate as Ezafe presence accuracy, and false positive rate as Ezafe absence accuracy. The total average can be calculated using a weighted average of the two last mentioned measures.

8 Experiments and Results

As mentioned, the system was trained on Bijankhan corpus. Only the first level of POS tags was used for the training phase, except for the words with Ezafe, that the POS plus Ezafe marker was chosen. The more fine-grained POS tags were removed to achieve more accuracy.

We used a ten-fold cross-validation scheme. For calculating the total accuracy, Ezafe presence accuracy and Ezafe absence accuracy should be weighted by 16.8% (ratio of words with Ezafe marker in test corpus) and 83.2% (ratio of words without Ezafe marker) respectively.

A. Evaluating fine-grained tags

The first experiment was done in order to test the ability of other fine grained POS tags in Ezafe recognition. In this test that was done on 30% of the corpus, all of the fine grained POS tags of the words plus Ezafe marker were used to train a Maximum Entropy POSE tagger. As shown in Table 2, the accuracy of the system decreased when we used complete features hierarchy. So, in consequent experiments, we used only first level tag features.

Conditions	**Performance measures (Run on 30% of corpus)**				
	Precision	***Recall***	***F-measure***	***MCC***	***Accuracy***
MaxEnt+ POSE	*87.95*	*93.14*	*0.91*	*0.89*	*96.71*
MaxEnt+ POSE+ fine grained tags	*89.56*	*88.69*	*0.89*	*0.87*	*96.37*

Table 2: Experiment Based on Full Tag Hierarchy

B. Evaluating MaxEnt tagger

In the next experiment, we used a MaxEnt tagger applied on whole corpus. With first level hierarchy of POSE tags, a total accuracy of 97.21% was resulted. As shown in the Table 3, while we have a good recall rate, the precision

reached a fair value. Both F-measure and MCC have values greater than 0.9.

The effect of eliminating titles which are incomplete sentences was also experimented. Table 3 shows that eliminating the titlesdoes not achieve a good improvement in accuracy.

C. *Using Persian-specific features*

Augmenting the system with some Persian-specific features to decrease FP and FN can significantly increase the total accuracy. As shown in Table 3, by using five features, the accuracy can be increased by more than 0.6%. The features are as follow:

- Punctuations cannot take Ezafe. By this simple feature, these FP errors will be removed.
- Noun words which are followed by adjectives and adverbs should take Ezafe marker.
- Adjectives which are followed by nouns and adverbs should take Ezafe marker.
- Adverbs which are followed by nouns and adverbs should take Ezafe marker.
- Nouns, adverbs and adjectives which are followed by verbs do not take Ezafe.

Conditions	**Performance measures (%)**				
	Precision	***Recall***	***F-measure***	***MCC***	***Accuracy***
MAXent+POSE	*89.44*	*94.48*	*0.919*	*0.903*	*97.21*
MAXent+POSE without title	*89.53*	*94.47*	*0.919*	*0.903*	*97.23*
Maxent+POSE+ Persian Specific Features	*91.37*	*95.92*	*0.936*	*0.923*	*97.80*

Table 3 - Results of Experiments on complete corpus Size

Note that the false positive rate of the above mentioned experiment is about twice of the false negative rate. So, we tried to extract more features based on investigating words in FP table and confusion matrix.

D. *Evaluating CRF Tagger*

The next experiment was based on CRF tagger. In order to compare the results with MaxEnt tagger, the experiment was performed on whole corpus using 10-fold cross validation method.

In this experiment, we used a CRF tagger and applied a window on the text to see the effect of neighboring words as input features in Ezafe recognition. As shown in Figure 1, the accuracy of system varies by changing the size of the window from 1 to 9. The graph shows that the experiments with a CRF tagger can achieve its best accuracy with window of size 5. Better performance was achieved by augmenting the CRF model with the five mentioned Persian-specific features.

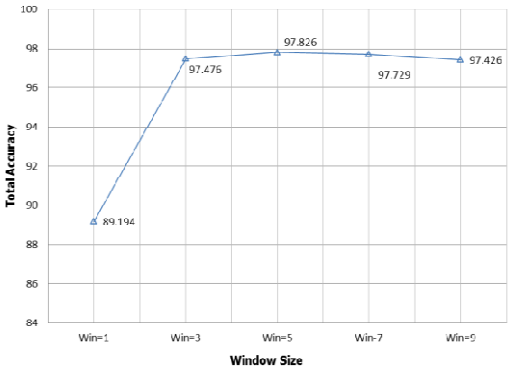

Fig. 1. Ezafe Recognition Accuracy vs. Window Size

Table 4 shows the results comparing with previous works in this regard. As shown in the table, the accuracy of CRF model augmented by mentioned featuresets can achieve best results comparing to other approaches.

Conditions	Ezafe presence accuracy	Ezafe Presence Error	Ezafe Absence Accuracy	Ezafe Absence Error	Total Accuracy
Rule based and syntactic (Oskouipour, 2011)	*10.37*	*89.63*	*83.20*	*16.80*	*70.06*
PCFG with 1000 sentences (Isapour, 2007)	*86.74*	*13.26*	*95.62*	*4.38*	*93.29*
Pattern based method patterns with freq>1% (Bijankhan, 2005)	*79.69*	*20.31*	*92.95*	*7.05*	*89.86*
HMM with 3gram (Oskouipour, 2011)	*78.55*	*21.45*	*95.31*	*4.68*	*91.69*
SMT based approach	*75.96*	*24.05*	*89.99*	*10.01*	*88.86*
MaxEnt with POSE	*94.48*	*5.52*	*97.75*	*2.25*	*97.21*
MaxEnt with POSE + Persian Specific Features	*95.92*	*4.08*	*98.18*	*1.82*	*97.80*
CRF Winsize=5	*95.15*	*4.85*	*98.36*	*1.63*	*97.83*
CRF Winsize=5 +Persian Specific Features	*96.42*	*3.58*	*98.367*	*1.63*	***98.04***

Table 4 - Comparison of results (%)

9 Conclusions

In this paper, we proposed a POSE tagging approach to recognize Ezafe in Persian sentences. Besides to this probabilistic approach, some features were extracted to increase the recognition accuracy. Experimental results show that CRF tagger acts pretty well in Persian Ezafe recognition. The obtained results show outstanding performance comparing to earlier approaches and the accuracy is quite reliable because of training based on a 10 million-words corpus.Future research can be done based on other taggers such as log-linear and TnT taggers. Moreover, Ezafe recognition can be viewed as a spell checking problem.So, a spell checker can also be used as another approach.

References

Amiri, Hadi, Hojjat, Hossein, and Oroumchian, Farhad.,2007.*Investigation on a Feasible Corpus for Persian POS Tagging*. 12th international CSI computer conference, Iran.

Bijankhan, Mahmoud.,*The Role of the Corpus in Writing a Grammar: An Introduction to a Software*, Iranian Journal of Linguistics, vol. 19, no. 2, fall and winter 2004.

Bijankhan, Mahmoud., 2005. *A feasibility study on Ezafe Domain Analysis based on pattern matching method*. Published by Research Institute on Culture, Art, and Communication, Tehran, Iran.

Estaji, Azam., Jahangiri, Nader., 2006 *The origin of kasre ezafe in persian language.* Journal of Persian language and literature, Vol. 47, pp 69-82, Isfahan University, Iran.

Farghaly, Ali., 2004. *Computer Processing of Arabic Script-based Languages: Current State and Future Directions.* Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004, University of Geneva, Geneva, Switzerland, August 28, 2004.

Ghomeshi, Jila. 1996. *Projection and Inflection: A Study of Persian Phrase Structure.* PhD. Thesis, Graduate Department of Linguistics, University of Toronto.

Isapour, Shahriyar., Homayounpour, Mohammad Mehdi, and Bijankhan, Mahmoud., 2007. *Identification of ezafe location in Persian language with Probabilistic Context Free Grammar*, 13th Computer association Conference, Kish Island, Iran.

Kahnemuyipour, Arsalan., 2003. *Syntactic categories and Persian stress.* Natural Language & Linguistic Theory 21.2: 333-379.

Lafferty, John.,McCallum, Andrew., and Pereira. Fernando, C.N., 2001 *Conditional random fields: Probabilistic models for segmenting and labeling sequence data,* In Proc. of ICML, pp.282-289,

Mavvaji, Vahid., and Eslami, Moharram., 2012. *Converting persian text to phoneme stream based on a syntactic analyser.* The first internationasl conference on persian text and speech, September 5,6, 2012, Semnan, Iran.

Namnabat, Majid., and Homayounpour, Mohamad Mehdi., 2006. *Text to phoneme conversion in Persian language using multi-layer perceptron neural network*, Iranian Journal of electrical and computer engineering, Vol. 5, No. 3, Autumn 2007.

Oskouipour, Navid., 2011. *Converting Text to phoneme stream with the ability to recognizing ezafe marker and homographs applied to Persian speech synthesis.* Msc. Thesis, Sharif University of Technology, Iran.

Pantcheva, Marina Blagoeva., 2006. *Persian Preposition Classes.* Nordlyd; Volume 33 (1). ISSN 0332-7531.s 1 - 25.

Parsafar Parviz. 2010, *Syntax, Morphology, and Semantics of Ezafe.*Iranian Studies[serial online]. December 2010;43(5):637-666. Available in Academic Search Complete, Ipswich, MA.

Razi, Behnam, and Eshqi, Mohammad, 2012. *Design of a POS tagger for Persian speech based on Neural Networks*, 20th conference on Electrical Engineering,15-17 may 2012,Tehran, Iran.

Samvelian, Pollet. 2007. *The Ezafe as a head-marking inflectional affix: Evidence from Persian and Kurmanji Kurdish.* Aspects of Iranian Linguistics: Papers in Honor of Mohammad Reza Bateni, 339-361.

Sha, Fei., and Pereira, Fernando, 2003. *Shallow parsing with conditional random fields,* In Proc. of HLT/NAACL 2003.

Shakeri, Zakieh, et al. 2012. *Use of linguistic features for improving English-Persian SMT*. Konvens 2012, The 11th Conference on Natural Language Processing, Vienna, Sept 19-21, 2012

Toutanova, Kristina Klein, and Manning. ChristopherD., 2000. *Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger*. Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

Toutanova, Kristina Klein, Manning, Christopher D., and Singer, Yoram. 2003. *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network.* In Proceedings of HLT-NAACL 2003, pp. 252-259.

Yarowsky, David. 1996. *Homograph disambiguation in text-to-speech synthesis; Progress in Speech Synthesis.* eds. van Santen, J., Sproat, R., Olive, J. and Hirschberg, J : 157-172.

Zahedi, Morteza., 1998. *Design and implementation of an intelligent program for recognizing short vowels in Persian text*. Msc. Thesis, University of Tehran.