Ezafe Prediction in Phrases of Farsi Using CART

Abbas Koochari [a][1], Behrang QasemiZadeh [b], Mojtaba Kasaeiyan [a]

[a] *Amirkabir University, Tehran, Iran*
[b] *Iran University of Science and Technology, Narmak, Tehran, Iran*

Persian, also known as Farsi, is the official language of Iran and Tajikistan and one of the two main languages Spoken in Afghanistan. Persians adopted a unified Arabic script for writing. In consequence, short vowels usually are not written in Farsi text. On the other hand, *Ezafe* marker, the genitive marker of Farsi, usually appears as a short vowel in Farsi written texts and it is not written in texts; so, Farsi written text, is a series of consecutive nouns without any overt links or boundaries. This can be lead to complexity and inefficiency of Farsi Syntactic analysis. The paper introduces a corpus based method to detect *Ezafe* marker in Farsi written texts to overcome the mentioned problem. Classification and Regression Tree (CART) has been used to predict the presence or absence of *Ezafe* marker. Evaluation shows promising results of our approach.

Keywords: Persian, Natural Language Processing, CART, Corpus

1 INTRODUCTION

Farsi, also known as Persian, is the official language of Iran, Tajikistan and one of the two main languages spoken in Afghanistan. Farsi is a member of the Indo-Iranian family of the Indo-European languages and it has the properties of agglutinative languages. [1][2]

After the Arab's conquest in 651 A.D., the Persians adopted an extension of unified Arabic script for writing [3]. Salient characteristics of Arabic script are: existence of various connecting letters, varying graphic forms for many letters depending on their position in a word, varying letter width, absence of full size characters for vowels (vowels are represented as particular signs above and below characters), existence of a number of digraphs and composite letters, writing direction from right to left and absence of upper case and lower case letters. As Farsi uses an extension of Arabic writing system, short vowels are not written in Farsi texts. [4]

In Farsi, the *Ezafe* Marker, a suffix that connects the elements in a phrase, may accompany nouns and adjectives. *Ezafe* Marker usually appears as a short vowel named *Kasre*, which sounds "e". *Ezafe* Marker acts like "`s" (e.g., John's car) or the preposition "of" (her brother's car) in English. A noun that is accompanied by *Ezafe* Marker, can be considered as the genitive case of that noun. According to the Farsi orthography, Due to the fact that *Ezafe* Marker appears as a short vowel, it is possible that it is omitted from written text. The result, in Farsi written text, is a series of consecutive nouns without any overt links or boundaries as shown in the following example [5]:

Māshīn dūst brādr Ali
Car friend brother Ali
Ali's brother's friend's car

The actual pronunciation for this example in spoken language is "Māshīn-*e* dūst-*e* barādar-*e* Ali", where the *Ezafe* marker is represented by the –*e*. [5]

One of the important issues about *Ezafe* marker is that all elements occurring between the head noun and the possessor noun phrase are linked to the head noun and to one another by *Ezafe* marker. In addition, within an adjectival phrase, *Ezafe* may link the adjectival head to its unique complement. There are several studies about *Ezafe* in Farsi from linguistics point of view. Samiian [6] is the first detailed study on Ezafe in Persian within a modern syntactic framework, namely X-bar theory. The empirical facts mentioned by Samiian have been taken

[1] Corresponding Author: Abbas Koochari, a_koochari@yahoo.com

up in subsequent works [7][8], although they have been accounted for in a radically different way.

In this paper we have proposed a corpus based method for detection *Ezafe* marker in Farsi written text. The proposed method uses Classification and Regression Tree (CART) to predict present or absence of *Ezafe* marker in Farsi written text. The paper is organized as follow: next section describes the corpus briefly, description of CART can be found in section 3. Experimental result and evaluation of method are discussed in section 4. Finally, we conclude in section 5.

2 THE CORPUS

Farsi version of 1984 corpus in MULTEXT-East framework [9] has been used for train and test the proposed method. The corpus approximately comprises of hundred thousand of words. Farsi version of corpus is annotated systematically according to the special PoS categorization of MULTEXT-East framework. According to their PoS categorization, there are 11 groups of words with their special attributes. Table 1 shows the PoSs and the Number of their attributes.

Table 1. Farsi PoSs according to MULTEXT-East framework.

Part of Speech	Code	Number of Attributes
Noun	N	4
Verb	V	10
Adjective	A	4
Pronoun	P	6
Determiner	D	1
Adverb	R	2
Adposition	S	2
Conjunction	C	2
Numeral	M	1
Interjection	I	0
Abbreviation	Y	0

In the proposed framework in [9], they indicate *Ezafe* as the genitive case marker for nouns and adjectives of Farsi words. More detail about the attributes can be found in [9] and [10].

3 CART: CLASSIFICATION AND REGRESSION TREE

CART methodology was developed in 80s by Breiman, Freidman, Olshen, Stone in their paper "Classification and Regression Trees” [11]. CART is a statistical modeling technique used to predict a value of a variable y using the corresponding feature vector f. CART is a binary branching tree with questions about the influencing factors at the nodes and predicted values at the leaves. A CART-based modeling successively divides the feature space to minimize the prediction error. Finally, it constructs a tree representing the partition of the feature space.

CART analysis is a form of binary recursive partitioning. Each node is split into two child nodes, in which case the original node is called a parent node. The term recursive refers to the fact that the binary partitioning process is applied repeatedly to reach a given number of splits. In order to find the best possible split features, all possible splits are calculated, as well as all possible return values to be used in a split node. The program seeks to maximize the average ``purity" of the two child nodes using the misclassification error measure. [12]

4 THE EXPERIMENT

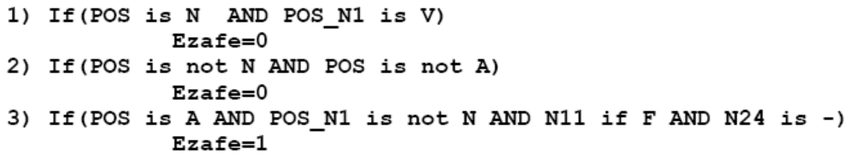
In our case study, binary classification trees were trained to predict the presence or absence of the *Ezafe* marker between two adjacent words of the training corpus. Since there is no previous knowledge about usefulness of the features and their relative importance, CART's are built in a step-wise method to construct the usefulness and relative importance of the features. In this approach, each single feature is taken in turn and a tree consisting of nodes containing only the conditions imposed by that feature is built. The single best tree is then kept and each remaining feature is taken in turn and added to the tree to find the best possible tree with just two features. The procedure is then repeated for the third, fourth, fifth feature and so on. This process continues until no significant gain in accuracy is obtained by adding more features. Edinburgh Speech Tools Library [17] has been used to build CART from the corpus.

Only morphosyntactic features of words have been used for training and construction of tree. For this reason, a window with 6 neighbour words, 2 words before and 3 other words after the current word, were selected. The features that were used to train the system for current word comprises of PoS of current word and its neighbors as well as their corresponding morphosyntactic features.

Table 2. The experimental results

Type of Prediction	**Total Num.**	**Num. of Correct Predictions**	**Score**
non-*Ezafe*	27494	27015	%98.25
Ezafe	2878	2557	%88.85
Total (combined)	30372	29572	%97.366

The corpus was divided to 2 different part, one for training and the other for the test. The training part consisted of approximately 70000 words. Remaining of the corpus were used as the test set. Table 2 shows the result of our experiment. The test corpus for evaluating the model was comprised of 30372 words where 2878 of them were accompanied by the Ezafe marker. In the case of non-Ezafe words, system prediction was right for 88.85% of cases. As non-Ezafe words, system prediction came true in 98.25%. Figure 1 shows number of rules which are extracted from CART.



```
1) If(POS is N  AND POS_N1 is V)
            Ezafe=0
2) If(POS is not N AND POS is not A)
            Ezafe=0
3) If(POS is A AND POS_N1 is not N AND N11 if F AND N24 is -)
            Ezafe=1
```

Fig. 1. Examples of rules that are extracted from CART

Precise evaluation could be shown by the Kappa factor. This measure was first suggested for linguistic classification tasks by Carletta [13] and has since been used by others to avoid the dependency of the score on the proportion of non-breaks in the text. The kappa statistic is calculated as indicated by following Formula:

$$k = \frac{\Pr(A) - \Pr(E)}{1 - \Pr(E)} \tag{1}$$

Where Pr(A) is the overall score attained by the tree and Pr(E) is the proportion of non-

Ezafe words in the corpus.

In expression, overall score achieved by the tree is compared with the probability of having a non-*Ezafe* label in the corpus, eliminating the dependency on the structure of the data. If the algorithm does not insert any *Ezafe* marker, the value of the kappa statistic will be 0. If the method predicts every inter-word *Ezafe* marker correctly, then k=1. Values lower than 0 indicate that the *Ezafe* markers placed by the algorithm are in the wrong places.

Regarding the Formula (1) and the experimental result shown in table 2, Pr(A) and Pr(E) are 0.97366 and 0.9052 respectively. With these values, the kappa factor is 0.72.

5 CONCLUSION

This Paper introduces a corpus-based method for detecting the *Ezafe* marker in written texts of Farsi. *Ezafe* Marker is a suffix that connects the elements in a phrase. It usually appear as a short vowel named *Kasre* which sounds "e" and it is possible that this marker omitted from written text according to the orthography of Farsi. The *Ezafe* Marker acts like "`s" (e.g., John's car) or the preposition "of" (her brother's car) in English. Our method uses CART to model the absence or presence of *Ezafe* Marker. Farsi version of 1984 corpus in MULTEXT-East framework has been used for training the model.

Evaluation of system shows promising results to solve the problem. One of the important issues about the proposed method is the accuracy and consistency of the corpus that is used to train the model with the nature of problem. The corpus annotation in 1984, classifies Farsi words in different categories that are suitable to our approach. It is obvious that changes in annotation of corpus effects in the results of the model. In addition, the annotation system has a great influence on the efficiency of the system. The kappa factor for our experiment is 0.72.

REFERENCES

[1] Kalbasi I., The Derivational Structure of Word In Modern Farsi, ISBN 964-426-128-3, Tehran, Iran, 2001.

[2] Samare I., Typological Features Of Farsi, Journal Of Linguistics, Iran University Press, No. 7, pp 61-80,1990.

[3] Fischer S. R., History Of Writing, Reaktion Books, ISBN: 1861891679, 2001.

[4] Challenges Persian analysis

[5] Hassel M. & Mazdak N., FarsiSum - A Persian text summarizer. In the proceedings of Computational Approaches to Arabic Script-based Languages, Workshop at Coling 2004, the 20th International Conference on Computational Linguistics, Geneva, Switzerland, 2004.

[6] Samiian V., Origins of Phrasal Categories in Persian : An X-bar Analysis, UCLA, 1983.

[7] Kahnemuyipour A. 2000. Persian Ezafe Construction Revisited: evidence for Modifier Phrase. Paper presented at Proceedings of the Canadian Linguistic Association Conference, Edmonton.

[8] Ghomeshi G. 1997. Non-Projecting Nouns and the Ezafe Construction in Persian. Natural Language and Linguistic Theory 15-4:729-788.

[9] QasemiZadeh B. & Rahimi S., Persian in MULTEXT-East Framework, 5th intenational Conference on natural language processing, Springer-Verlag, Turku, Finland, 2006.

[10] Erjavec T., MULTEXT-East Morphosyntactic Specifications, Version 3.0. Supported By EU Projects Multext-East, Concede And TELRI, 2004.

[11] Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and regression trees, Monterey, Calif., U.S.A.: Wadsworth, Inc, 1984.

[12] Mitri, S. Frintrop, S. Pervolz, K. Surmann, H. Nuchter, A., Robust Object Detection at Regions of Interest with an Application in Ball Recognition, Proceedings of the 2005 IEEE International Conference on Robotics and Automation, ICRA, 2005.

[13] Carletta J. C., Assessing agreement on classification tasks: the kappa statistic, Computational Linguistics, 22(2), 249-254, 1996.