IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011
ISSN (Online): 1694-0814
www.IJCSI.org

Recognizing Anaphora Reference in Persian Sentences

Farshid Fallahi and Mehrnoush Shamsfard

Department of Computer Engineering, Shahid Beheshti University

Tehran, 19839-63113, Iran

Abstract

Finding the reference of pronouns in a piece of text, which is a type of co-reference resolution, is an important task in discourse analysis and processing natural language texts.Pronoun reference is the noun that is replaced by the pronoun.In this paper, we propose a rule-based method for pronoun reference resolution in Persian texts. Our method exploits rules to recognize the reference of various types of pronouns in a 3-sentences interval. An automatic reference resolution system is developed based on the proposed method as the first pronoun reference resolution system for the Persian language. The experimental results show admissible accuracy in test cases. In this paper firstly we will describe some problems and challenges in detecting pronoun references and have an overview of related works in this field. In the next sections, after a brief description of the proposed method and the developed system, its features and architecture, we will discuss its components in detail. Then we will explain the experimental results and discuss further works to improve the system.

Keywords:*Pronoun Reference, Reference Resolution, Discourse Analysis, Natural Language Processing, Persian.*

1.Introduction

Co-reference Resolution refers to recognizing the reference of various entities such as pronouns and proper nouns in a piece of text. In other words it talks about determining which noun phrase is referring to which real world entity mentioned in the text. Co-reference resolution and its variants such as anaphor resolution or pronoun reference resolution usually work outside of a single sentence and so are counted as discourse analysis tasks in natural language processing. They are important tasks in many NLP applications such as machine translation, text understanding, questionanswering, text summarization, and so forth.

Pronoun (anaphora) reference is the noun that is replaced by the pronoun and usually appears somewhere before it. Sometimes a pronoun has no reference; i.e. it appears as a noun, or its reference does not appear directly in the text. Usually, however, pronoun reference appears before the pronoun, and sometimes after it.

Thus pronouns usually refer to other words, called their antecedents because they (should) come before the pronoun. A pronoun's antecedent may be either a noun or another pronoun, but in either case, it *must*be clear what the antecedent is. The term pronoun reference or anaphora reference describes the relationship between the pronoun and its antecedent. Problems occur when the pronoun's antecedent is unclear or ambiguous. In such cases readers (human or machine) have problem deciding which of two or more earlier nouns a pronoun stands in for, or have difficulty finding any noun at all.

There are some differences in anaphora system between Persian and English which cause the need to some new methods for this language.Persian is a null-subject, or pro-drop language, so personal pronouns (e.g. I, he, and she) are optional. Pronouns generally are the same for all cases including nominative, accusative and possessive.The first-person singular accusative pronoun has two regular and short forms (*mænrā* "me" can be shortened to *mærā)*. Table1 shows the normal form of Persian pronouns.

Table 1: Normal form of Persian pronouns

Normal detached Forms		
Person	Singular	Plural
1st	mænمن	māما
2nd	to تو	shomāشما
3rd	u او (non-human/human), vey وی (human only and formal)	ānhāآن‌ها (non-human/human), ishānایشان (human only and formal)

Possession can be expressed either by normal forms of pronouns or by adding suffixes (genitive enclitics) to nouns. These are added after inflection for number (table 2).

Table 2:Genitive enclitics

Genitive enclitics		
Person	Singular	Plural
1st	-æm	-emān
2nd	-æt	-etān

3rd	-æsh	-eshān

Note that when the stem to which these are added ends in a vowel, a "y" is inserted for ease of pronunciation.

There is also another type of accusative pronouns which can be used in conjunction with verbs to incorporate a direct object (table 3).

Table 3:Attached accusative pronouns

Attached accusative pronouns		
Person	Singular	Plural
1st	æm	emān
2nd	æt	etān
3rd	æsh	eshān

This type usually is used in colloquial speech while using it in written formal texts is not prohibited. For example the translation of the sentence "Yesterday I bought it." in two forms with detached and with attached accusative pronoun are shown in table 4.

Table 4:Example for Direct object incorporation

Example for Direct object incorporation	
diruzanrakharidam	دیروز آن را خریدم
diruzkharidæmæš	دیروز خریدمش

At last there are 7 reflexive pronouns in Persian; 6 of them for 6 persons (like for English) and the last is person-less pronoun which can be used instead of all 6 persons and numbers. Table 5 shows the list of these pronouns.

Table 5: Example for Direct object incorporation

Reflexive pronouns		
Person	Singular	Plural
1st	Khodæm (myself)	Khodemān (ourselves)
2nd	Khodæt (yourself)	Khodetān (yourselves)
3rd	Khodæsh (him/her/it self)	Khodeshān (themselves)
ALL	khod	

Persian pronouns are neutral and do not have gender information. In addition there are some exceptions in number agreement between the pronoun and its antecedent.We sometimes use singular pronouns to refer to inanimate plural antecedents and sometimes use plural pronouns to refer to singular antecedents (for respecting). These exceptions cause difficulties in pronoun reference resolution.

In general three types of problems in recognizing anaphora reference are:

- Ambiguous Reference:A pronoun reference is ambiguous if it has more than one alternative and the reader cannot easily understand which of the earlier nouns is its real antecedent.
- Vague or Indefinite Reference: In some cases we use pronouns it or they without mentioning the reference explicitly in the text. In these cases we use these pronouns to refer to vague, unnamed entities.
- Implied Antecedents: Sometimes the pronoun antecedent is not a noun but a fact implied in a phrase, sentence or a piece of text. In such we wish to refer to the whole idea of the statement.

In this paper, we propose a rule-based method for pronoun reference resolution in Persian texts. Our method exploits some rules to recognize the reference of various types of pronouns in a 3-sentences interval. An automatic reference resolution system is developed based on the proposed method. Experimental results show admissible accuracy in test cases which outperforms the other available work.

In the rest of the paper, firstly we will describe some problems and challenges in detecting pronoun referencesand have an overview on the related work in this field. In the next sections, after a brief description of the proposed method and the developed system, its features and architecture, we will discuss its components in detail. Then we will explain the experimental results and discuss further works to improve the system.

2.Related Work

Anaphora reference resolution may be done by rule based methods or machine learning approaches. Rule based methods exploits some rules to determine the reference of pronouns. The rules are usually defined manually. On the other side, machine learning algorithms need a training set of texts annotated by pronoun references. This set may be provided manually or semi-automatically. In the manual case, the user first manually labels anaphora reference in a set of training texts. A learning system then generates rules from the training texts. These rules can then be employed to recognize the reference of pronouns from new texts. The main problem with machine learning methods is the lack of training sets in less-studied languages such as Persian.

In 2007, Culottaet al.[1] proposed a machine learning method that enables features over sets of noun phrases.They outline a set of approximations that make this approach practical, and apply this method to the ACE co-reference dataset, achieving a 45% error reduction over a comparable method that only considers features of pairs of noun phrases.Denis and Baldridge[2] proposed a supervised ranking approach for pronoun resolution. The ranking enables all candidate antecedents to be evaluated together; whereas classification methods examine at most two candidate antecedents at a time.

They showed that their proposed method outperforms the bestclassification method.
In 2004, XiaoqiangLuo and Abe Ittycheriah[3] proposed a new approach for co-reference resolution which uses the *Bell tree* to represent the search space and casts the co-reference resolution problem as finding the best path from the root of the Bell tree to the leaf nodes. A Maximum Entropy model used to rank these paths.

In 2002, Ng and Cardie[4] presented a noun phrase co-reference system that extends the work of Soon et al. (2001). Improvements arise from two sources: extra-linguistic changes to the learning framework and a large-scale expansion of the feature set to include more sophisticated linguistic knowledge.

There is also a work related to Persian anaphora resolution. In 2009, Sadat Moosavi and Ghassem-Sani[5] investigated approaches to Persian pronoun resolution. They have tested some known methods in the field of classification, machine learning, and ranking on a small set of 90 manually tagged documents randomly taken from Peykareh corpusBijankhan. The results are not very promising.

In most of machine learning works it is assumed that there is a training set to make the system applicable. While there is no such a set available for Persian language and creating a large training set manually is time and cost consuming, in this paper, we propose a rule based method for pronoun reference resolution. Although it is not a complex system, it can be used to develop a training set for a machine learning anaphora resolution system.

In fact, there is no previous works on rule-based Persian coreference/pronoun resolution. Our proposed rule-based method for Persian coreference resolution has high accuracy and in some cases acts better than machine learning algorithms. We evaluate our proposed method and describe some empirical evaluation in experimental results section in this paper.

3. The Proposed Method

In this section we propose a rule based method to determine the pronoun references in a window of three sentences in a Persian text. The architecture of our developed system for Persian Pronoun reference resolution is shown in fig.1. As the fig.1 shows the operational system consists of two main parts; preprocessing and resolution. In preprocessing phase the input text is processed to extract a sequence of POS tagged tokens. To do this we firstly use a tokenizer to determine the word boundaries in the sentence. Then using a POS tagger [6] we assign a POS tag to each word in the sentence. The small tag set we used in our system is shown in table 6.After preprocessing the input will be fed into the anaphora resolution subsystem to find the reference of pronouns according to some manually built heuristic rules.

The last part is the evaluation module whichevaluates the system's performance and decides about changing the rules if needed.

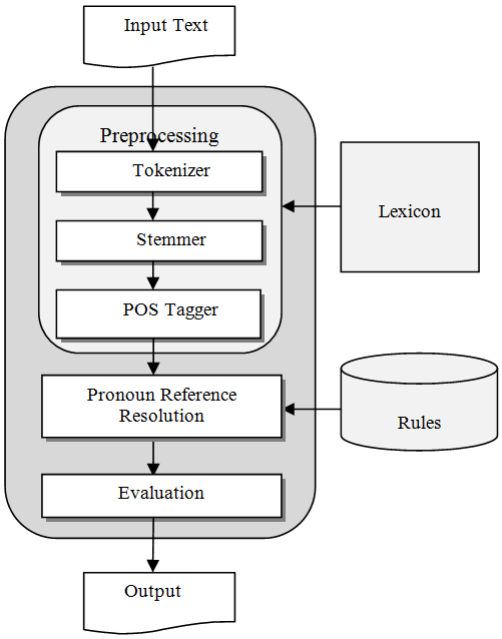

Fig. 1 The architecture of our developed system for Persian Pronoun reference resolution.

Table 6:The small tag set we used in our system

AR (Arabic)	PUNC (Punctuation)
PRO (Pronoun)	N (Noun)
POSTP (Post preposition)	CONJ (Conjunction)
V (Verb)	N_PR (Proper noun)
ADV. (Adverb)	ADJ (Adjective)
P (Preposition)	N_CN (Counting number)

By investigating various Persian sentences, we extract some heuristic rules based on which we determine that our input text is similar to which of the rules. Then, we recognize the pronoun reference.

The main part of our project is devoted to recognizing accusative pronoun references so that we could determine these references with high accuracy using the

extracted rules. As it was mentioned, first, we receive a Persian statement in input, and then after tokenization and stemming, wetag all of the statement words using POS Tagger. Words tagged in PRO tag are the pronouns whose reference should be determined using manual rules.

In this section we show some instances of our heuristic rules which are extractedfor recognizing nominative anaphora references:

I. If the pronoun is in the nominative part of second sentence, then antecedent is more probable to be found in the nominative part of first sentence.

II. If the pronoun is in the objective part of second sentence, then its reference is more probable to be found in objective part of first sentence.

III. If we have pronouns "it" and "he(she)" in a sentence, then pronoun reference of "he(she)" will be a noun that is animate and pronoun reference of "it" is a noun that is inanimate noun in the previous sentence.
Example:
John saw the book. He bought it. (“Johnketabra did. Ouanrakharid.”)
According to the above rule, the pronoun reference of “he” (“ou”) is John and the pronoun reference of “it” (“an”) is book.

IV. If the pronoun is "they", its antecedent is either a plural noun or some nouns attached by ‘va’ (and)
Example:
Ali and Reza have an exam today. They were not satisfied with the test.
(“Ali va Rezaemroozdaremtehansherkatkardand.Anhaazemtehanrazinabudand”.)
According to the above rule, the pronoun reference of “they”(“anha”) is “Ali and Reza” (“Ali va Reza”).

V. If we have pronoun "we" in a sentence, pronoun reference will be “Iand” with the word after it.
Example:
I and Reza went to Hussein's home. We bought a gift for him.
(“manvareza be khaneyehosseinraftim. mabarayeouhediyekharidebudim.”)
According to the above rule, the pronoun reference of “we”(“ma”) is "I and Reza" (“man vareza”).

VI. If the pronoun is "there" or “here”, its reference should be a location or place in previous sentences. The referring place may appear as a location adverb or a noun in a propositional phrase in previous sentences. It may have been tagged by location in previous stages.

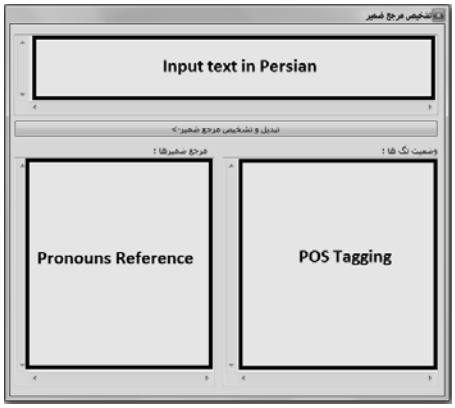

Fig. 2 a view of our program.

We use some other rules that are similar to those above. Our high accuracy in recognizing pronoun (anaphora) reference in Persian sentences is due to our accurate rules. We evaluate our proposed method and describe some empirical evaluation in experimental results part in this paper.

In order to find accusative pronoun reference in Persian sentences, we first use a stemmer which determines the attached accusative pronouns (such as “his” or “her” (“sh”), “my”(“am”), etc.) in input Persian text. Then using some heuristic rules similar to the above-mentioned rules, we recognize objective pronoun reference with high accuracy.
Example:
Farshid saw Ali. He took his book. (“Farshid Ali ra did. Ketabashragereft”)
Using manually extracted rules for recognizing objective pronoun reference, our program displays “his” (“sh”) pronoun reference in the output.

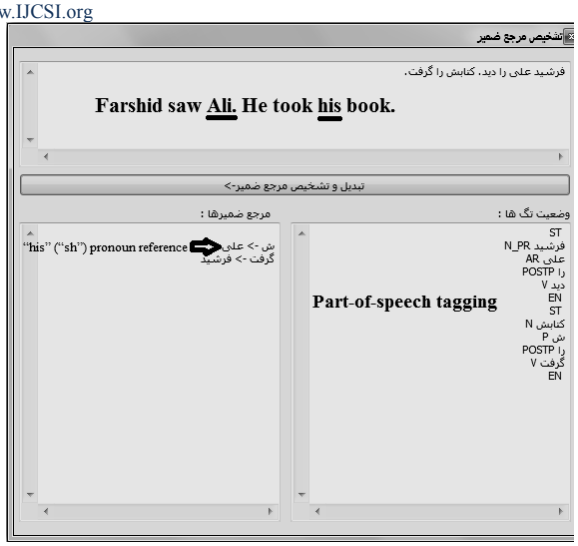

Fig. 3 an Example.

Analyzing the source of errors show that the following factors decreases the performance of the system:

- **Intrinsic ambiguities and Errors in natural language**
  Ambiguity is a natural characteristic of natural languages. Something is ambiguous when it can be understood in two or more possible senses or ways. The ambiguity may occur in word level (such as lexical ambiguity), in sentence level (such as structural ambiguity or word sense ambiguity) or in discourse level (such as pronoun reference ambiguity). In the last case even humans may not easily find the reference of a pronoun among the alternatives.

- **Errors in preprocessing**
  POS tagger's Accuracy that we used in the preprocessing part is below 95 percent and thus it can cause someerrors in assigning POS tag to each word in the sentence. Wrong POS tagging will result in wrong referenceresolution as the rules are highly dependent to words' POS tags.

- **Exceptions in heuristic rules**
  Although we tried to build the heuristic rules as general as possible, in some cases, a rule may result in awrong output in a specific sentence and this can be a source of error.
  In the next section we will show the experimental results and compare our work with the other system available for Persian.

4. Experimental Results

This project is done in ShahidBeheshti University. Based on the proposed approach, we built a program using C# for recognizing pronoun reference in Persian sentences. We now evaluate our proposed method, and compare it with the machine learning method proposed in [5]. The experimental results are given in Table 1. Below, we first describe some experimental settings and then discuss the results.

Test Texts: Five Persian blogger websites are used in our experiments. From each Website, 20 randomweblog pages are downloaded.

Evaluation measures: We use the standard precision and recall measures to evaluate the results of system.
Table 7 shows the results for the experiment. In this table 'miss' shows the number of cases in which the text contains the pronoun reference, but it is not found and 'wrong' shows the number of cases in which the page has the pronoun reference, but a wrong one is found.

Table 7: Experiment results

	Blogger Website	No. of random pages that is used for test	No. of pronoun reference that the text from 20 pages contains it (sum of 20 pages)	Our proposed method (Rule-based method)		Machine Learning method	
				miss	wrong	miss	wrong
1	Mihanblog.com	20	257	13	12	9	15
2	Persianblog.ir	20	102	5	7	12	8
3	Blogfa.com	20	69	3	8	4	10
4	Iranblog.com	20	196	11	5	10	6
5	Parsiblog.com	20	119	8	2	14	3
	Total	100	743	74		91	

Table 7 shows the precision and recall of applying two methods (our proposed rule based method and the machine learning method proposed by [5]) on the provided test bed. Results show that our method outperforms the only other system available for Persian anaphora resolution.

Table 8:comparing our proposed method with the other available system for Persian

	Recall	Precision
Our Rule-based Method :	90%	95.1%
Machine Learning Method :	87.7%	92.7%

5. Conclusions and Future Works

This paper proposed a rule-based approach to anaphora reference resolution for Persian texts. Unlike existing methods, the proposed method does not perform machine

learning to generate rules based on a set of user-labeled training pages. Our algorithm can start anaphora reference resolution from a set of manual extracted rules and either annotate the pronouns by their references or develop a training set for machine learning approaches.

At present, the accuracy of our program is over 90% in Persian weblogs texts, as described in the previous part. This accuracy is due to the employed rules. If we increase the number of rules, the accuracy of the program increases. We plan to do this in our future work, especially increasing the objective pronoun reference rules.Combining our method with a machine learning approach to increase the performance is also among our further works. Experimental results with pronoun reference extraction from 100 Persian Weblogs pages show that our proposed approach is highly effective.

References

[1] AronCulotta, MichaelWick, Robert Hall, Andrew McCallum, "First-Order Probabilistic Models for Coreference Resolution", In Proceedings of HLT-NAACL, 2007.

[2] Pascal Denis, Jason Baldridge, "Specialized models and ranking for coreference resolution", In Proceedings of the Conference on Empirical Methods in Natural Language Processing (2008), pp. 660-669.

[3] XiaoqiangLuo, Abe Ittycheriah, " A Mention-Synchronous Coreference Resolution Algorithm Based on the Bell Tree", In Proceedings of the ACL, 2004.

[4] Vincent Ng, Claire Cardie, "Improving Machine Learning Approaches to Coreference Resolution", In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2002.

[5]Nafiseh Sadat Moosavi, Gholamreza Ghassem-Sani, "A Ranking Approach to Persian Pronoun Resolution", 10th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2009), Mexico City, Mexico, 2009.

[6] MehrnoushShamsfard, HakimehFadaee, "A Hybrid Morphology-Based POS Tagger for Persian". In Proceedings of 6th Language Resources and Evaluation Conference (LREC 2008), Morocco, 2008.