Design and Implementation of Anaphora Resolution in Punjabi Language

Kawaljit Kaur[1], Vishal Goyal[2], Kamlesh Dutta[3]
[1,2]Department of Computer Science, Punjabi University, Patiala, India
[3]Department of Computer Science and Engineering, NIT, Hamirpur, HP,India
[1]saini_kawal@rediffmail.com,[2]vishal.pup@gmail.com,[3]kdnith@gmail.com

Abstract

Natural Language Processing (NLP) is the most attention-grabbing field of artificial intelligence. It focuses on the interaction between humans and computers. Through NLP we can make thec omputers recognize, decode and deduce the meaning ofhuman dialect splendidly. But there are numerous difficulties that are experienced in NLP and, Anaphora is one such issue. Anaphora emerges often in composed writings and oral talk. Anaphora Resolution is the process of finding antecedent of corresponding referent and is required in different applications of NLP.Appreciable works have been accounted for anaphora in English and different languages, but no work has been done in Punjabi Language. Through this paper we are enumerating the introduction of Anaphora Resolution in Punjabi language. The accuracy achieved for the system is 47%.

1 Introduction

Humans have an incredible capability to interact or communicate with one another. Language acts as a powerful medium for this communication. It helps people to express their thoughts using different words but it is quite complex. Complexity of the language can be reflected from the fact that same thing can be narrated in a different no. of ways and a sentence can be interpreted by various people in distinct ways.

A crucial element of the language is the occurrence of the reference which is the process of using language representation to select an "entity" in the real world. The "entity" can be an object in the physical world or a concept in our mind. The linguistic symbol is termed as "referring expression". The issue of reference is the problem of establishing a relationship between different parts of discourse to understand the content which is of great importance for a computational linguistic.

The importance of the problem of reference can be understood from its real time applications in NLP which are - Question Answering/Information Extraction, Automatic Summarization, and Machine Translation as represented in (Aalan F. Smeaten,1994).

ANAPHORA resolution is one of the most sophisticated and complicated problem in modern language processing. The problem of research work is to put focus on Punjabi language in contrast of ANAPHORA RESOLUTION and have fine understanding of the interaction between the syntax and semantics of the language. There are various types of anaphora but pronouns as anaphora are more frequently encountered. So, focus will be on resolving anaphors that act as pronouns.

2 Main Components Of Anaphora Resolution System

The main components of Anaphora Resolution System are:

2.1 Standard pre-processing tools

At the initial stage, the text is required to be converted into the form so that it can be processed by the system. Punjabi Shallow Parser has been used during the pre-processing stage, which performs Morphological analysis and POS tagging. The output is in SSF. Corpus is manually annotated with attributes given in Table-1:

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 35–36
Patna, India, December 18 - 21, 2020.

S.No	Attribute	Used with	Explanation
1	name	All chunks	Gives identification to all phrases within the sentence e.g if it is first NP in the sentence then it will be given name as NP, next NP found in the sentence will be given name= 'NP2' and so on.Similary for other phrases.
2	ref	Pronouns/anaphoras	This establishes anaphoric link.It contains name of the phrase which acts as antecedent for the anaphora.If corresponding antecedent (e.g' NP1')is in same sentence then ref= 'NP1' .But if antecedent is in different sentence then ref='..%2%NP1' i.e. pronoun refers to NP1 of sentence no 2.
3	refType	Pronouns/anaphoras	It specifies the type of anaphora. It can take value 'C' for concrete anaphora or 'E' for event/abstract anaphora.
4	dir	Pronouns/anaphoras	It specifies the direction of antecedent in the discourse. It can take value "A" if antecedent/referent is anaphora(backward direction. It can have value 'C' if antecedent/referent is cataphora (forward direction)
5	PType	Pronoun Type	It specifies type of pronoun: PER3,PER2,PER1 …Third Person/Second Person/First Person REF….Reflexive REL….Relative
6	Animacy	Antecedent	Human,Animate,Inanimate
7	NER	Antecedent	PERSON,ORG,LOC,FAC

Table 1: Annotation Attributes

2.2 Task specific pre-processing module (NP finder and Pronoun Finder)

It deals with extracting Noun Phrase and Pronouns from the sentence. It takes input as sequence of parsedtokens, identifies NP's and discards rest of the tokens. It is used during the step of finding the anaphora and its possible antecedents.

2.3 Feature Extraction Module

Machine Learning Approach has been used as computational strategy and feature extraction is a very important task. It represents data as feature vector of attributes and value pairs. The features describe the properties of anaphora, its antecedent candidates or their relationship. The features that have been extracted from text are- Agreement Feature ,Distance Features, Animacy, NER , Pronoun Type and Referrent Type(Dakwale et al,2012).These features are then applied on model for classification.

2.4 Classification Method

Classification Methods are used to predict most appropriate antecedent for the anaphora. It takes input from Feature Extraction module, performs processing based on set of rules and find the unambiguous antecedent for the anaphora. The classification method which has been employed is Naïve Bayes Classifier.

3 Results

Presently, the system has been checked on a single file having 238 pronouns. The classifier used is Naïve Bayes Classifier. Efficiency of the system is measured using Recall and F score.
Recall:0.46 ,F-Score: 0.63
Overall Accuracy: 47%
Similar results were found during initial work in Hindi language also.

References

Aalan F. Smeaten, Progress in the Application of Natural Language Processing to Information Retrieval Tasks, The Computer Journal, Volume 35, Issue 3, 1992, pp 268-278.

Dakwale, Praveen, Himanshu Sharma, and Dipti Misra Sharma. "Anaphora annotation in hindi dependency treebank." In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, 2012, pp. 391-400.