International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-4, June 2015
ISSN: 2395-3470
www.ijseas.com

A Machine Learning Approach to Resolve Event Anaphora

Komal Mehla[1], Ajay Jangra[1], Karambir[1]
[1] University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra, India

Abstract
Anaphora Resolution is considered as one of the important area in the field of Natural Language Processing. A lot of research has been done in anaphora resolution for English Language, but in Hindi language the research is limited. Most of the researches for Hindi anaphora resolution work for entity anaphora resolution. This paper presents a machine learning approach which can works for event anaphora resolution in Hindi Language. An algorithm is described for resolving event anaphora and the accuracy of the algorithm is also checked for different algorithms.

Keywords: *Natural language Processing (NLP), Anaphora Resolution (AR), Hindi shallow Parser (HSP), Noun Phrase (NP), Part of Speech (POS), Verb Group Finite (VGF), Verb Group Infinite (VGNN).*

1. Introduction

In linguistics, anaphora is the use of an expression which refers to another expression in its context. The referring expression itself is called anaphor and referred expression is called the antecedent (referent). The antecedent provides the information for the interpretation of the anaphor. Consider a simple example:

(1) Director of XYZ bank ordered an investigation against the regional manager. He claimed that the manager is involved in recently discovered loan scams.

In the above example, 'He' is referring to a previously mentioned entity 'Director of XYZ bank'. The process of identification of the referent is known as 'Anaphora Resolution' the term used for reference to an expression occurring later than the pronoun is 'cataphora'. However, in computational linguistics, the usage of the term 'anaphora' stands for the reference to any expression which may come earlier or after the anaphor. While there has been significant research for anaphora resolution in English and other languages, NLP research including anaphora resolution for Indian languages is limited.

One categorization of anaphora can be event and entity anaphora. The reason for having separate resolution process for Event and Entity anaphora is obvious. In concrete or Entity reference, an anaphora is referring to a concrete entity, thus possible referents are Noun phrases, while in Abstract or event reference, an anaphor refers to an abstract object, thus the possible referents are verbs, clauses and propositions. Moreover, features and linguistic properties of candidates in both cases are different. Thus it is efficient to consider separate resolution processes for both types of references [1]. In Hindi (and in many other languages), for some pronouns, it is difficult to determine whether they refer to an entity or event, only on the basis of their lexical form, hence in order to consider resolution of both the types separately, a classification module is required a priori which classifies ambiguous anaphors into Entity and Event references.

2. Related work

Finley and Joachims [3] in 2005 described an approach to supervised clustering. The algorithm learnt a similarity measure to produce desired clustering. This contrasts with using pairwise classification, where the target concept to be learned is "same cluster or not". The sparsity of such pairs in the training data (only around 1.6% of the pairs in MUC-6 training set are co-referent) made the data set imbalanced. However, Ng and Cardie [2] used only the closest preceding non-pronominal noun phrase for a given phrase to make a positive pair, and all co-referent pairs between the two were paired to form negative samples. Further, like the CRF approach this algorithm too could take care of transitive dependencies. The main difference with the CRF approach was that the objective to be maximized is the margin instead of the likelihood as in the CRF method. One problem was that the number of constraints in this formulation can grow faster than exponential with the number of items. Suitably

optimizing the objective function was an NP-complete problem, and therefore approximate inference methods were adopted. One of the advantages is that this method can handle transitive dependencies; however, when there is not much transitive dependency in the dataset, the performance of this method is only comparable to that of pairwise classification. Further, it was not clear whether this approach was better than just correlation clustering or the CRF approach discussed earlier.

A. Bharti ***et al.*** [4] in 2006, worked with a purpose to arrive at standard tagging scheme for POS tagging and chunking for annotating Indian languages (AnnCorra) and came up with the tags which are exhaustive for the task of annotation for a larger group of languages, specially, Indian languages. The document gave a detailed description of the tags which had been defined for the tagging schemes and elaborates the motivations behind the selection of these tags. The document also discussed various issues that were addressed while preparing the tag sets and how they had been resolved.

Sachin Aggarwal *et al.* [5] in 2007, presented anaphora resolution as a technique of semantic analysis of text documents written in Hindi language. In this technique semantically related sentences in a text were located through anaphors. Their semantics are analyzed and exploited for resolving referents of the respective anaphors. Matching constraints for the grammatical attributes of different words was the basis of that approach. The accuracy of anaphora resolution was 96% for simple sentences and nearly 80% for compound and complex sentences.

K. Dutta et al. [6] in 2008 worked for pronominal resolution in Hindi, for which they studied various different applications of Hobbs algorithm. Hobb's algorithm is used as baseline algorithm because it uses syntactic information instead of semantic information. For Chinese, to resolve third person pronouns 77.6% accuracy is obtained. Since Spanish and Turkish are free-word order languages hence there are difficulties in implementing Hobb's algorithm. Hence the algorithm has been modified in order to use for these languages. Full parsing was not used while implementing Spanish and noun phrases are searched from left to right to adjust specifications. After that it is tested for those including in NP (Noun Phrase) and an interrupt is generated when NP agreed for number and gender with the anaphor. The algorithm report accuracy of 62.7%. After slight modification and using full parsing the algorithm can deal with pronominal possessive and null pronouns for Turkish language. Since Hindi is also free word order language, hence the antecedents for reflexive and possessive pronouns can be found similarly.

B. Uppalapu and D. M. Sharma [7] in 2009, presented an algorithm which was in S-List [1] (Prasad and Stube, 2000) for resolving the Hindi third person pronouns. They showed that there is an improvement in the performance of the S- List algorithm when the discourse entities of the present utterance and the entities of the previous utterances were considered into two different lists rather than using single list to consider all the entities. The antecedents for the first and second person pronouns can be found with the help of verbs and their modifiers. They explored how complex sentences can be broken into utterances. For resolving the first and second person pronouns a new approach was also introduced by them.

Zhang Ning ***et al.*** **in** [8] in 2011, described an machine learning based framework for event pronoun resolution. Both flat and structural features are explored for event pronoun resolution in this system. Minimum Expansion Tree (MET) and Semantic Role Expansion Tree (SRET) were two structural features which were examined. They contained different substructures of the parse tree. Convolution tree kernel was used to compute the similarity between two structural features. The kernel enumerates all the sub trees of given two trees and number of common sub-trees is used as the measure of similarity between two trees. Random down sampling method was used during training to do the sampling of positive and negative instances.

K. Karthikeyan et al. [9] in 2013 presented a novel enhanced framework for Anaphora resolution process. Proposed system is able to recognize Pronoun phrase, Intra and Inter sentinel anaphora, Animacy Agreement and co-reference chain. The system recognized maximum anaphora so that it produces more accuracy antecedents compared to prior anaphora resolution process models.

Donghong Liu [10] in 2013 pointed out that discourse entity cannot be used in event anaphora if it is considered as discourse topic; that if a discourse topic takes the form of question, anaphora resolution may not be realized because of the uncertainty in the question and in the components of the question; and that if a discourse topic takes null form, the global

coherence and the relevance of a discourse might be undermined. The paper also proves that comparatively propositional discourse topics conform to human beings' cognitive psychology, contribute to event anaphora resolution and facilitate discourse construction.

S. Lalitha Devi et al. [11] in 2014 presented anaphora resolution system for all Indian languages. Here the authors have described the system architecture and its implementation. The authors have chosen two Indian languages representing two major families of languages of India. They are Hindi and Tamil belonging to Indo-Aryan and Dravidian families respectively. The system is specifically designed to be scalable and allows plug-n-play architecture. The core anaphora resolution engine uses CRFs a machine learning technique. This uses feature based learning and have provided syntactic and positional based features and obtained encouraging evaluation results. The evaluation metrics used are standard measures MUC, B-CUBED and CEAF and obtained an average of 53.85% F-measure for Hindi and 53.95% of F-measure for Tamil.

S. Lalitha Devi et al. [12] in 2014 described the first effort on automatic identification of connectives and their arguments for three Indian languages Hindi, Malayalam and Tamil. They have adopted machine learning technique Conditional Random Fields (CRFs) for their work and used a corpus of 3000 sentences belonging to health domain. Domain independent features were extracted to improve the performance of the system. They mainly concentrated on the identification of explicit connectives and their arguments. Two sets of experiments were performed. First set of experiment was performed for the identification of connectives and next for the identification of argument boundaries. Using this approach, they obtained encouraging results for all the three languages. Error analysis shows the presence of different structural patterns of discourse relations among three languages.

Srishti Singh et al. [13] in 2014 presented paper talks about the application of the Bureau of Indian Standards (BIS) scheme for one of the most widely spoken Indian languages 'Bhojpuri'. Bhojpuri has claimed for its inclusion in the Eighth Schedule of the Indian Constitution, where currently 22 major Indian languages are already enlisted. Recently through Indian government initiatives these scheduled languages have received the attention from Computational aspect, but unfortunately this non-scheduled language still lacks such attention for its development in the field of NLP. The present work is possibly the first of its kind. The BIS tagset is an Indian standard designed for tagging almost all the Indian languages. Annotated corpora in Bhojpuri and the simplified annotation guideline to this tagset will serve as an important tool for such well-known NLP tasks as POS- Tagger, Phrase Chunker, Parser, Structural Transfer, Word Sense Disambiguation (WSD), etc.

We have also done some work related to anaphora resolution. In this work, we have discussed some background of event anaphora resolution and compare the different approach of anaphora resolution [14].

3. Proposed Algorithm

We resolve the event anaphora using the machine learning approach. The various steps used are below:

- Parse the sentence using Hindi shallow parser.
- Considered the parsed text of current sentence and previous one.
- Extracted all 'VGF' and 'VGNN' chunks : candidates (on an average 5 candidates per anaphora)
- Classified each independently.
- Produced result based on classifier output for all the candidates.

But system is difficult to learn because the number of positive references given to classifier is less than the number of negative references given to classifier. So, the system will be more biased towards negative class. For example, consider a chance based system with 5 candidates for an anaphora. Then, the following result are obtained.

Probability of getting correct output for candidate=0.5

Probability of getting wrong output for candidate=0.5

Probability of overall correct output= 0.5*0.5*0.5*0.5*0.5= 0.03125~ 3.12%

2.1 The training and test data

One third (1/3rd) of the data is held out (fixed) for testing. Thus the testing data contains 1071 concrete pronouns. The remaining two-third (2/3rd) data is

used in 4-fold iteration for training and development, i.e. the remaining data is again divided randomly in 4 parts and evaluated in 4 iterations. In each iteration, one fourth of data is used as development (tuning) and the remaining three fourth (3/4th) is used for training (in supervised learning). For the rule based setting the only parameter which we tune is the number of sentences 'n' (that should be considered while searching for the referent).

Note that since pronouns occur in discourse, training and testing set are divided in files and not in pronouns. Thus across all the iteration, it is the number of files that remain same but the number of pronouns may vary since texts (files) may contain different number of pronouns.

We have tested our approach on different occurrence of event pronouns in sentences. In some cases, it works perfect but in some cases, inappropriate results are obtained. Consider an example:

इसके साथ साथ संयुक्त घोषणा जारी करने पर भी विचार किया जा रहा है | हालाँकि माकपा को इस पर कुछ एतराज है|

The shallow parsed text is below:

4 ((VGNN <fs name='VGNN' drel='k7:VGF'>

4.1 करनेVM <fs af='कर,v,any,any,any,o,ना,nA' name='करने' posn='80'>

4.2 पर PSP <fs af='पर ,psp,,,,,,' name='पर ' posn='90'>

4.3 भी RP <fs af='भी,avy,,,,,,' name='भी' posn='100'>

))

. . .

3 ((NP <fs name='NP2' drel='k7:VGF' semprop='rest'>

3.1 इस PRP <fs af='यह,pn,any,sg,3,o,,' name='इस' posn='40' ref='VGNN' refmod='NP2/JJP'reftype='E'>

3.2 पर PSP <fs af= 'पर,psp,,,,,,' name='पर' posn='50'>

))

3 ((NP <fs name='NP2' drel='k7:VGF' semprop='rest'>

3.1 इस PRP <fs af='यह,pn,any,sg,3,o,,' name='इस' posn='40'ref='VGNN'refmod='NP2/JJP' reftype='E' **predRef='VGNN'**>

3.2 पर PSP <fs af='पर,psp,' name=पर ' posn='50'>

))

In the above example, the proposed approach works fine. The anaphora इस is linked to the event करने पर भी correctly.

4. Results

We have tested our algorithm on different machine learning algorithms. The result values obtained are shown in the following table 1.

Table 1: Results

Algorithm	Accuracy
Bagging	46.15%
CART	47.55%
Random Forest	48.95%
Decision Tree	51.74%
Random Subspace	53.84%
Decorate of decision trees	58.04%

5. Conclusions

This paper described a machine learning algorithm for event anaphora resolution. It is noticed that there is difficulty while learning due to large difference in the number of positive references and negative references. The algorithm works very well for large number of sentences, but in certain cases confusing results are obtained. However, the accuracy of the algorithm on different algorithms is quite impressive. In future the algorithm can be enhanced by

considering more features. As a result of which the accuracy of system for different algorithms can be improved.

References

[1] R. Prasad and M. Strube, "Discourse salience and pronoun resolution in Hindi", U. Penn Working Papers in Linguistics, 6 pages 189-208, 2000.

[2] V. Ng and C. Cardie, "Improving machine learning approaches to co-reference resolution", In ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA , pages 104–111, 2001.

[3] T. Finley and T. Joachims, "Supervised clustering with support vector machines", In ICML '05: Proceedings of the 22nd international conference on Machine learning, New York, NY, USA, pages 217–224, 2005.

[4] A. Bharati, D. M. Sharma, L. Bai, and R. Sangal, "Anncorra: Annotating corpora guidelines for pos and chunk annotation for Indian languages", Technical report, LTRC, IIIT-Hyderabad, 2006.

[5] S. Agarwal, M. Srivastava, P. Agarwal, and R. Sanyal, "Anaphora resolution in Hindi documents", In Natural Language Processing and Knowledge Engineering, 2007(NLP-KE 2007) International Conference , pages 452–458. IEEE, 2007.

[6] K. Dutta, N. Prakash, and S. Kaushik, "Resolving pronominal anaphora in Hindi using Hobbs algorithm", Web Journal of Formal Computation and Cognitive Linguistics, 1(10), 2008.

[7] B. Uppalapu and D. M. Sharma, "Pronoun resolution for Hindi", In 7th Discourse Anaphora and Anaphor Resolution Colloquium, 2009.

[8] Zhang Ning, Kong Fang, Li Peifeng, "Research of Event Pronoun Resolution", In IEEE International Conference on Asian Language Processing, 2011.

[9]K.Karthikeyan and Dr.V.Karthikeyani, "Understanding Text using Anaphora Resolution", In Proceedings of the IEEE International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME) February 21-22, 2013.

[10] Donghong Liu, "Discourse topic in Anaphora Resolution and Discourse Construction", In Proceedings of the IEEE International Conference on Asian Language Processing, 2013.

[11] Sobha Lalitha Devi, Vijay Sundar Ram and Pattabhi RK Rao, "Anaphora Resolution System for Indian Languages", 9th International Conference on Language Resources and Evaluation, in LREC 2014.

[12] Sobha Lalitha Devi, Sindhuja Gopalan and Lakshmi S, "Automatic Identification of Discourse Relations in Indian Languages", 9th International Conference on Language Resources and Evaluation, in LREC 2014.

[13] Srishti Singh and Esha Banerjee, "Annotating Bhojpuri Corpus using BIS Scheme", 9th International Conference on Language Resources and Evaluation, in LREC 2014.

[14] K. Mehla, Karambir and A. Jangra, "Event Anaphora resolution in natural language processing for Hindi text", in International journal of innovative science, engineering and technology(IJISET), vol. 2, issue 1, January 2015