A Generic Anaphora Resolution Engine for Indian Languages

Sobha Lalitha Devi
AU-KBC Research Centre
MIT Campus of Anna University, Chennai, India
sobha@au-kbc.org

Vijay Sundar Ram
AU-KBC Research Centre
MIT Campus of Anna University, Chennai, India
sundar@au-kbc.org

Pattabhi RK Rao
AU-KBC Research Centre
MIT Campus of Anna University, Chennai, India
pattabhi@au-kbc.org

Abstract

In this paper, we present a generic anaphora engine for Indian languages, which are mostly resource-poor languages. We have analysed the similarities and variations between pronouns and their agreement with antecedents in Indian languages. The generic algorithm developed uses the morphological richness of Indian languages. The machine learning approach uses the features which can handle major Indian languages. We have tested the system with Indo-Aryan and Dravidian languages namely Bengali, Hindi and Tamil. The results are encouraging.

1 Introduction

Natural language has different types of anaphoric expressions and these expressions bring elegance and make the natural language text interesting to read. Anaphoric expression in a discourse refers to another item in a discourse. The task of resolving anaphors with its referent, antecedent is called as anaphora resolution. Anaphora resolution is required in most of the NLP applications to achieve required performance. The importance of anaphora resolution in various tasks is demonstrated by researchers by integrating anaphora resolution with answer extraction system, automatic summarization system, relation extraction system, document similarity identifier etc.

Most of the anaphora resolution systems are developed for particular languages. The researchers have analyzed anaphors across languages at various levels such as syntactic, semantic, discourse, structured, and unstructured features. But there are very few attempts for language independent approaches. In this paper, we present a generic anaphora resolution engine for Indian languages. We have come up with a language independent engine, which takes shallow parsed text as input. The morphological richness of Indian languages is tapped to come up with a language independent anaphora resolution engine.

Early works in anaphora resolution by Hobbs (1978), Carbonell and Brown (1988), Rich and LuperFoy (1988) etc. were mentioned as knowledge intensive approach, where syntactic, semantic information, world knowledge and case frames were used. Centering theory, a discourse based approach for anaphora resolution was presented by Grosz (1977), Joshi and Kuhn (1979), Joshi and Weinstein (1981), Strube and Hahn (1999). Salience feature based approaches were presented by Lappin and Leass (1994), Kennedy Boguraev (1996) and Sobha et al., (2000). Indicator based resolution methods were presented by Mitkov (1997, 1998). One of the early works using machine learning technique was Dagan Itai's (1990) unsupervised approach based on co-occurrence words. With the use of machine learning techniques researchers work on anaphora resolution and noun phrase anaphora resolution simultaneously. The other machine learning approaches for anaphora resolution were the following. Aone and Bennett (1995), McCarty and Lahnert (1995), Soon et al., (2001), Ng and Cardia (2002) had used decision tree based classifier. Daelman and Van de Bosh (2005), Hendrickx et al., (2008), Recasen (2009) had used TiMBL, a memory based learning approach. Anaphora resolution using CRFs was presented by McCallum and Wellner (2003) for English, Li et al., (2008) for Chinese and Sobha

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1824–1833, Dublin, Ireland, August 23-29 2014.

et al., (2011, 2013) for English and Tamil. Expectation Maximization (EM) was used for anaphora resolution by Charniak and Elsner (2009). Wich et al., (2012) demonstrated a coreference resolution in a large scale using discriminative hierarchical model.

In Indian languages anaphora resolution engines are demonstrated only in few languages such as Hindi, Bengali, Tamil, and Malayalam. Most of the Indian languages do not have parser and other sophisticated pre-processing tools. The earliest work in Indian language, Vasisth was a rule based multilingual anaphora resolution platform by Sobha and Patnaik (2000, 2002), where the authors had exploited the morphological richness of Malayalam and Hindi. Prasad and Strube (2000), Uppalapu et al., (2009) and Dekwale et al., (2013) had presented different approaches using Centering theory for Hindi. Dutta et al (2008) had presented a Hindi anaphora resolution system using Hobbs' algorithm. Murthy et al., (2007) had presented a comparison on Tamil anaphora resolution using multi-linear regression and salience factor based approach. Sobha et al., (2007) presented a salience factor based with limited shallow parsing of text. Akilandeswari et al., (2013) used CRFs for resolution of third person pronoun and Akilandeswari et al., (2012) presented a work on resolution of 'atu', third person neuter pronoun in Tamil. Balaji et al., (2012) presented resolution using two stage bootstrapping approach. The author had used UNL representation. Ram et al., (2013) used Tree CRFs for anaphora resolution for Tamil with features from dependency parsed text.

One of the earliest multilingual anaphora resolution systems was presented by Aone and Mckee (1993), where the authors had used Global discourse world which contained syntactic, semantic, rhetorical and other information. They demonstrated the system for English, Spanish and Japanese. Mitkov (1998) extended the indicators based approach for other languages and presented it for English, Polish and Arabic. As mentioned earlier, Vasisth was the only multilingual attempt for Indian language anaphora. SemEval-2010 Task 1: Coreference resolution in Multiple Language, a tool contest accelerated the research in multilingual anaphora resolution. The contest had six languages, namely, English, German, Italian, Dutch, Spanish and Catalan. There were six participants, among them only two participants presented results for all six languages. The systems presented in the contest were RelaxCor, Corry, SUCRE, BART, TANL-1 and UBIU (Recasens et al., 2010). The multilingual anaphora resolution in Indian languages was re-initiated by Anaphora Resolution in Indian languages, the tool contest conducted as a part of ICON 2011 (Sobha et al., 2011). The contest had three languages, namely, Hindi, Bengali and Tamil. There were four participants and all the four submitted results for Bengali. Tamil and Hindi had two participants. This task boosted the anaphora resolution work in Bengali. Senapati et al., (2013) presented a work on Bengali anaphora resolution by customizing GuiTAR and BART tool was customized for Bengali by Sikdar et al., (2013). In all the above published multilingual systems, there was a language dependent module plugged in. In the present work, we have tried to come up with an approach without using language specific modules. We have developed a generic anaphora engine as the system is designed to work for different languages. We have overcome the agreement problem with the PNG information obtained from in-depth morphological analysis and PNG agreement heuristic rules. These rules are capable of filtering the possible candidate antecedents for an anaphor (pronouns) using the PNG information across languages.

The rest of the paper is organised as follows: In the following section we have described nature of Indian languages and variation in antecedent-anaphor agreement in Indian languages. Section 3, we have explained our approach towards generic engine by overcoming the variation in antecedent-anaphor agreements. In section 4, we have presented our experiment and results. And the last is the conclusion section.

2 Characteristics of Indian Language Anaphora

Indian languages are morphologically rich and verb-final languages. These languages have relatively free word order and clausal structures are more fixed order. Indian languages fall under the following broader families of languages, Indo-Aryan, Dravidian and Tibeto-Burman. Indo-Aryan family includes languages such as Hindi, Bengali, Marathi, Punjabi etc, Dravidian includes languages such as Telugu, Kannada, Malayalam, Tamil, Tibeto-Burman includes languages such as Bodo, Manipuri etc. Dravidian languages are highly agglutinated and have rich productive suffixation than the Indo-Aryan languages. Plural marker and case markers get affixed to the nouns and tense markers and Person, Number, Gender (PNG) markers affix with verbs. In certain Indo-Aryan languages such as Hindi, case

markers occur as postpositions following the nouns. These postpositions are handled in the preprocessing stage to occur in the noun morphological analysis. Indian languages vary largely in the distinction of Number (singular/plural) and Gender in pronouns. Few of the Indian languages and their details of Number and Gender Distinction in those languages are presented in table 1.

Language	**Number Distinction (singular/plural)**	**Gender Distinction**
Hindi	Yes	No
Sanskrit	Yes	Yes
Punjabi	Yes	No
Gujarati	Yes	No
Assamese	Yes	No
Bengali	Yes	No distinction for Masculine and Feminine. But there is animate- inanimate distinction.
Oriya	Yes	No
Telugu	Yes	Masculine and others
Kannada	Yes	Yes
Malayalam	Yes	Yes
Tamil	Yes	Yes

Table 1: Variation of Pronouns with respect to Number and Gender

The similarities and variations between languages in the number-gender characteristics of the pronouns is presented in Table 1. In the number characteristics the languages are similar whereas in the gender characteristics there are variations. The information in the table 1 brings forth the challenges in capturing the anaphor-antecedent PNG agreement for a generic anaphora resolution engine.

With example 1, 2 and 3, we have demonstrated the variation in Gender, Number distinction in pronouns in Tamil (Ta), Bengali (Bn) and Hindi (Hi).

Example 1

```
Ta:a) raamum        giithavum      cakotharan-cakothari.
      Ram(N)+inc    Gita(N)+inc    brother   -sister.
      (Ram and Gita are brothers and sisters.)

   b) avan     elzhaam     vakuppu      padikkiraan.
      He(PN)   seventh(N)  standard(N)  study(V)+present+3sm
      (He studies in seventh standard.)

   c) aval     paththaam vakuppu      padikkiraal.
      she(PN)  tenth(N)  standard(N)  study(V)+present+3sf
      (She studies in tenth standard.)
```

Example 2

```
Bn:a) raam o giita bhai-bon.
   b) se shapton shreni te pore.
   c) se doshom shreni te pore.
```

Example 3

```
Hi:a) raam aur giitaa bhaii-bahan hai.
   b) vaha satavIM kakshaa meM paTataa hai.
   c) vaha aaTaviM kakshaa meM paTatii hai.
```

Example 1 has three Tamil sentences. The second and third sentence has pronouns 'avan' *he* and 'aval' *she*, the third person masculine and the third person feminine pronouns respectively. Masculine pronoun in second sentence refers to the masculine noun 'raam' and the feminine pronoun in the third sentence refers to the feminine noun 'Gita' in the first sentence. Here the masculine and feminine pro-

nouns have a clear distinction. The three Tamil sentences in example 1 are translated to Bengali and Hindi. The Bengali translation is presented in example 2 and Hindi translation is presented in example 3. In example 2, the second and third sentence has 'se', the third person pronoun. In the second sentence, the pronoun 'se' refers to the masculine noun 'raam' and 'se' in the third sentence refers to feminine noun 'giitaa' in the first sentence. Here the third person pronoun does not have masculine/feminine distinction. Similarly in example 3, which has the Hindi translation, 'vaha', the third person pronoun does not have gender distinction. In sentence 2 of example 3, 'vaha' refers to the masculine noun and in sentence 3, 'vaha' refers to the feminine noun.

These variations in Number and Gender distinction in pronouns pose challenges in coming up with a generic anaphoric engine. The pronoun and its agreement with its antecedents vary between the languages and to handle the agreement we require a language dependent mapping.

3 Generic Anaphora Resolution Engine

Most of the Indian languages are resource poor languages. The morphological richness of these languages, help in building various high end NLP applications such as machine translation, anaphora resolution etc., with limited shallow parsed information without using sophisticated parsing tools. In this work we have tried to build a generic anaphora resolution engine using shallow parsed text. Similarities between Indian languages, described in the previous section, are tapped to come up with a generic approach for anaphora resolution in Indian languages. The variation in the antecedent-anaphor agreement mentioned in the section above is handled by an in-depth morphological analysis of the text. We have used CRFs, a linear graphical machine learning algorithm to resolve the antecedents.

3.1 Preprocessing of Data

We perform limited shallow parsing on the training and testing data. Both the data are pre-processed with morphological analyzer, Part-of-Speech (POS) tagger, Chunker, Clause boundary identifier and Named Entity Recognizer. Here morphological analysis, Part-of-Speech tagging, Chunking are obligatory. Clause boundary identification and Named Entity Recognition are optional pre-processing tasks. These two tasks add information, which can be used as constraint features in the machine learning approach. In this work, we perform a detailed morphological analysis for a given word. This is explained in the following section. The preprocessing tools available in Indian Language –Indian Language Machine Translation (IL-ILMT) consortium are used.

3.2 Detailed Morphological Analysis

We perform an in-depth morphological analysis for a given word. In the in-depth morphological analysis we analyse both inflectional and derivational morphology. The in-depth morphological analysis gives the suffix (case markers with the nouns, tense-aspect-model with the verbs) and PNG characteristics of the words. These suffix information is used in the syntactic feature and verb suffix feature for the machine learning technique which are described further in section 3.4. The post-position occurring with the nouns, its syntactic association with the noun is identified in morphological processing stage and information is used as syntactic feature.

The morphological analyser identifies the root word, its lexical category, gender, number, person, case (direct/oblige), case markers if the word is a noun and tense markers (vibhakthi as called in Indian traditional grammar) if the word is a verb and the suffixes. Gender information holds information such as 'm' – masculine, 'f' – feminine, 'n' – neuter, 'mf' – can be a masculine or feminine, 'fn' – feminine or neuter as in Telugu and 'any' – can be any gender. Number information can be singular, plural, dual or any. Person information can be 1st person, 2nd person, 3rd person or any. We have explained it further with following example words and its analysed output in table 2.

S.No	Language	Word	Analysis of the Word
1	Ta	jaanukku 'John(N)+dative'	<fs af='jaan,n,m,sg,3,d,ukku,ukku'>
2	Ta	viittil 'house(N)+locative'	<fs af='viitu,n,any,sg,3,d,il,il'>
3	Ta	avanaal 'he(pn)+INS'	<fs af='avan,pn,m,sg,3,d,aal,aal'>
4	Ta	avalukku 'he(pn)+dative'	<fs af='aval,pn,f,sg,3,d,ukku,ukku'>
5	Hi	adhikaarii 'officer (N)'	<fs af='adhikaarii,n,m,sg,3,d,,'>

6	Hi	siitaa ‘sita (N)’	<fs af='siitaa,n,f,sg,3,d,,'>
7	Hi	uskaa ‘he/she/it (pn)’	<fs af='vaha,pn,any,sg,3,d,kaa,kaa'>
8	Hi	ve ‘they (pn)’	<fs af='vaha,pn,any,pl,3,d,,'>
9	Bn	chele ‘boy (N)’	<fs af=’chele,n,m,sg,3,d,,'>
10	Bn	meyze ‘girl (N)’	<fs af=’meyze,n,m,sg,3,d,,'>
11	Bn	se ‘he/she (PN)’	<fs af='se,pn,mf,sg,3,d,,'>

Table 2: Words and in-depth analysis

In the table 2, we have presented nouns and pronouns from Hindi (Hi), Bengali (Bn) and Tamil (Ta) and their in-depth analysis. The first word 'jaanukku' is a masculine singular noun. So the analysis has 'm,sg,3'. The second word 'viittil' is a neuter singular noun and its analysis has 'n,sg,3'. The third word is third person masculine and the fourth word is a feminine pronoun, so the analysis are ‘m,sg,3’ and ‘f,sg,3’ respectively. The words in the Fifth and sixth example are Hindi nouns with masculine and feminine gender respectively. The seventh example is third person singular and eighth example is third person plural word from Hindi. Hindi pronouns do not have gender distinction. The gender, number, person for the two pronouns are 'any,sg,3' and 'any,pl,3' respectively. 'any' in the gender slot shows the pronoun can refer to a noun phrase with any gender including neuter gender. The ninth and tenth example words are Bengali nouns ‘boy’ and ‘girl’, with masculine and feminine gender respectively. The eleventh word is a Bengali third person masculine and feminine pronoun. The gender, number, person in the morphological analysis has 'mf,sg,3'. This pronoun can refer to both masculine and feminine noun in Bengali.

3.3 Data Format

After pre-processing, the data is presented in a column format. The following are the columns information. First column has sentence id, followed by word id, POS tag, chunk tag, in-depth morphological analysis, clause information and Named Entity information. The training data has an additional column having the antecedent-anaphor agreement information.

3.4 Architecture of the Engine

The engine works independent of language. We have used heuristic rule based algorithm to select the candidate noun phrases for a given pronoun and machine learning techniques based approach to filter the exact antecedent noun phrase. As every supervised machine learning approach, this approach also has training and testing phase. The architecture of our approach for training and testing is given in figure1 and 2 respectively.

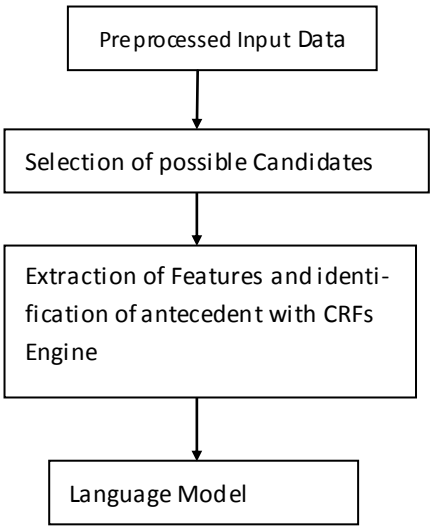

Figure 1: Training phase

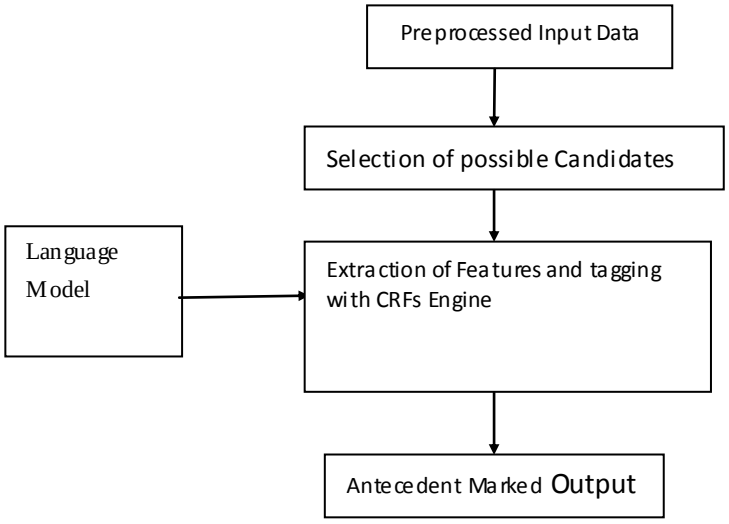

Figure 2: Testing phase

Selection of Candidate Noun Phrases for Antecedent

Both the training and testing phase has selection of the possible candidates. The noun phrases which agree with the pronoun in PNG should be selected as possible candidates for its antecedent. In training phase, the noun phrases which match with PNG of the pronoun and occur in between the anaphor and the antecedent are collected for each pronoun and given for training using the machine learning algorithm. The exact anaphor and antecedent pair forms positive pair and other noun phrases and anaphor form negative pairs for learning. In the testing phase all the noun phrases that match in PNG with the pronouns are collected from the current sentence and four prior sentences. The gender distinction and anaphor-antecedent agreement varies widely among Indian languages. In order to have a language independent engine, these variations have to be dynamically captured and the rules for checking PNG agreement have to be generated. We have used the gender information from the morphological analysis extracted with a set of heuristic rules to capture the variation in PNG agreement. The heuristic rules describe the possible genders that can match with the gender of the pronoun that varies between languages. The heuristic rules are presented below.

1. If the gender of the pronoun is 'm', then the nouns having masculine gender are chosen as candidate antecedents.
2. If the gender of the pronoun is 'f', then the nouns with feminine gender are chosen as candidate antecedents.
3. If the gender of the pronoun is 'n', then the nouns having neuter gender are chosen as candidate antecedents.
4. If the gender of the pronoun is 'mf', then the nouns with gender 'mf', 'm' and 'f' are chosen as candidate antecedents and the nouns with gender 'mf' is given importance.
5. If the gender of the pronoun is 'fn' is the gender of the pronoun, then nouns with 'fn' are chosen as candidate antecedents.
6. If the gender of the pronoun is 'any', then all the nouns are considered for candidate antecedent set and the nouns with gender 'any' is given higher priority.

Once the possible candidates for antecedents for the anaphors are selected, they are given for training/testing using the machine learning technique. Here we have used CRFs, a linear graphical technique to learn and identify the antecedents.

Anaphora Resolver

The core anaphora engine uses CRFs, a machine learning technique. In the training phase the system is provided with annotated data and the features for learning. After the system learns, a model file is generated as output. In the testing phase any unseen text is given for the automatic anaphora resolution. In our approach we have modeled this as a binary classification task. The machine has to classify whether the given candidate antecedent is the real antecedent or not based on the features of the candidate antecedents and the pronoun. The features for learning are extracted from the shallow parsed data. The feature extraction module extracts these features for all possible candidate antecedent and pronoun pairs from the shallow parsed data. The features used for learning are described below. Here we have used the freely available open source CRFs (Kudo, 2005).

Features for Learning

The features required for this task are identified from shallow parsed input sentences. The features for all possible candidate antecedent and pronoun pairs are obtained by preprocessing the input sentences with in-depth morphological analyser, POS tagger, and chunker, clause boundary identifier and Named Entity recognizer, where the last two preprocessing tasks are optional. The features identified can be classified as positional features, syntactic features and constraint features.

a) *Positional Features:* The occurrence of the candidate antecedent is noted. Is it in the same sentence where the pronoun occurs or in the prior sentences? Prior four sentences from the current sentence are considered.

b) *Syntactic Features:*

Syntactic Role: The syntactic role of the candidate noun phrases in the sentence is a key feature. The syntactic role of the noun phrases such as subject, object, indirect object, are obtained from the

case suffix affixed with the noun phrase. We consider Nominative and Dative cases for subject and other cases for object, the position and the other cases for indirect object.

Linguistic Characteristics: POS tag and chunk information of Candidate NP, suffixes affixed with the noun.

c) *Verb Suffixes:* The suffixes which show the gender which gets attached to the verb.

d) *Nature of NP:* Whether the candidate NP (probable antecedent) is Possessive or Existential.

e) *Constraint Features:* The constraint features are obtained from clause boundary and named entities recognized.

The position of the candidate NP with respect to clause boundary such as is candidate NP in current clause or immediate clause or non-immediate clause.

The Named Entity tags associated with the candidate NPs help the learning algorithm to learn constraints that types of NEs that can be its possible antecedents.

f) Combination of the above said features.

The features for learning have been identified based on the characteristics of the pronouns. For example constraint feature (current-clause) and syntactic feature (subject) helps in identifying the antecedent of the reflexives. For relative anaphors the constraint feature (immediate-clause) and syntactic feature (subject) help in identifying the antecedent.

4 Experiment, Results and Discussion

We have tested our approach using the dataset provided in "Anaphora Resolution for Indian Language", a tool contest conducted as a part of ICON 2011. The tool contest had three languages namely Tamil, Hindi and Bengali. The dataset is presented in column format with the following information viz. line index, word index, word, its POS and chunking information, followed by Named Entity information. We have enriched the dataset with in-depth morphological analysis. Table 3 presents the statistics of the ICON 2011 tool contest dataset.

Language	Training Data			Testing Data		
	Bengali	Hindi	Tamil	Bengali	Hindi	Tamil
Total Number of Pronouns	814	835	925	494	507	609
Number of Anaphoric Pronouns	476	557	580	283	344	348
Number of Non-Anaphoric Pronouns	338	278	345	211	163	261

Table 3: Statistics of ICON 2011 Dataset

MUC, B^3, BLANC, CEAF are the common scorers available for coreference resolution, where the coreference chains are being evaluated. Anaphora resolution is generally evaluated with performance measures such as Precision, Recall and F-measure. In this work, we have measured the performance with Precision, Recall and F-Measure and it is presented in table 4.

Example 4

Ta:	**aalluNar**	**rosaiyya**	villavil	kalanthukkoNtaar.	**avar**	athil	uraiyaRRinaal.
	Governor	Rosiah	function+loc	join+past+3SH .	He(gender neutral)	this	gave_lecture

(Governor Rosiah joined the function. He gave the lecture there.)

In above example 4, 'avar' (third honorific singular pronoun) in the second sentence, refers to a masculine noun phrase 'aalluNar rosaiyya' in the previous sentence. The honorific pronoun can also refer to a feminine pronoun. This possibility of referring to both the gender introduces more errors.

In Hindi, most of the pronouns such as "vaha" (he/she/it), "usa" (he/she/it), "unhone" (he/she honorofic) and "khuda" (himself) etc., do not have gender distinction and can be used to refer to antecedents of both feminine and masculine. PNG agreement adds more challenges in anaphora resolution, due to which the system gives more false positives. In our algorithm we were able to reduce the number of false positives and obtained better precision by having positional features and the verb suffixes as learning features. For languages such as Hindi, we observe that there is necessity of having verb

analysis in the text processing component. If we have this information as pre-processed data to the resolution engine, it can reduce ambiguity and improve the anaphora resolution.
As it is seen from the results table we have obtained lesser scores for Bengali. In Bengali third person pronouns such as “ami” (I), “túmi/tui/apni” (you), “se/tini” (he/she), “amra” (we), “tara/tnara” (they), do not have masculine, feminine distinction, but there is animacy distinction. And also the verb has no gender agreement. This adds more challenge to anaphora resolution engine and hence lesser scores than other languages. We identify the animacy feature in the morphological analysis stage for all the languages, but for Bengali language it is not robust and it affects in the anaphora resolution. For such languages the order of NPs and syntactic roles play major role in anaphora resolution. The use of features such as Named Entities helps in improving the resolution of pronouns referring to person and location. The addition of clause boundary information improves resolution by adding structural constraints.

5 Conclusion

We have presented a generic anaphora resolution engine, which can be used for all Indian languages. The three languages, Hindi, Bengali, and Tamil, we have chosen are the most spoken languages belonging to two major language families of India, namely belonging to Indo-Aryan and Dravidian families respectively. Though the Indian languages have similarities, they vary in Person, Number, Gender distinction, which pose a challenge in building language independent engine. The engine is language independent as it uses the information from the in-depth morphological analysis. It is specifically designed in such a way that it is scalable and allows plug-n-play architecture. The core anaphora resolution engine uses CRFs a machine learning technique. This uses feature based learning and we have provided syntactic and positional based features obtained from in-depth morphological analysis. We have obtained encouraging evaluation results. The major contributions of this work are the following:

a) Our attempt is first of its kind in Indian languages to develop a single generic engine, using machine learning.

b) It is a known fact that most of the Indian languages are resource poor, hence we have used very minimal resources, only shallow parsing has been used.

c) The results obtained are comparable to other reported works.

Reference

Akilandeswari A., and Sobha Lalitha Devi. (2013). Conditional Random Fields Based Pronominal Resolution in Tamil. *International Journal on Computer Science and Engineering*, Vol. 5 Issue 6 pp 601–610.

Akilandeswari A, Bakiyavathi T and Sobha Lalitha Devi, (2012), "atu Difficult Pronominal in Tamil", *In Proceedings of Lrec 2012*, Istanbul

Aone C., and McKee D. (1993).A Language-Independent Anaphora Resolution System for Understanding Multilingual Texts. *In proceeding of ACL 1993*, pp 156-163.

Aone C., and Bennett S. (1995). Evaluating automated and manual acquisition of anaphora resolution strategies. *In: 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 122-129.

Balaji J., Geetha T.V., Ranjani Parthasarathi R., Karky M. (2012).Two-Stage Bootstrapping for Anaphora Resolution *In: Proceedings of COLING 2012*, pp 507–516.

Carbonell J. G., and Brown R. D. (1988). Anaphora resolution: A multi-strategy approach. *In: 12th International Conference on Computational Linguistics*, 1988, pp. 96-101.

Charniak, E., and Elsner, M. (2009). EM Works for Pronoun Anaphora Resolution. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics* (EACL 2009), Athens, Greece.

Daelemans, W. and van den Bosch, A. (2005). Memory-based language processing. *Cambridge University Press*, Cambridge

Dagan I., and Itai. A. (1990). Automatic processing of large corpora for the resolution of anaphora references. *In: 13th conference on Computational linguistics*, Vol. 3, Helsinki, Finland, pp.330-332.

Dakwale. P., Mujadia. V., Sharma. D.M. (2013). A Hybrid Approach for Anaphora Resolution in Hindi. *In: Proc of International Joint Conference on Natural Language Processing*, Nagoya, Japan, pp.977–981.

Dutta. K., Prakash. N. and Kaushik. S. (2008). Resolving Pronominal Anaphora in Hindi using Hobbs‟ algorithm," *Web Journal of Formal Computation and Cognitive Linguistics*, Issue 10, 2008.

Li., F., Shi., S., Chen., Y., and Lv, X. (2008). Chinese Pronominal Anaphora Resolution Based on Conditional Random Fields. *In: International Conference on Computer Science and Software Engineering*, Washington, DC, USA, pp. 731-734.

Hendrickx I., Hoste V., and Daelemans W. (2008). Semantic and syntactic features for Dutch coreference resolution. In Gelbukh A. (Ed.), *CICLing-2008 conference*, Vol. 4919 LNCS, Berlin, Springer Verlag, pp. 731-734.

Hobbs J. (1978). Resolving pronoun references. *Lingua 44*, pp. 339-352.

Grosz, B. J. (1977). The representation and use of focus in dialogue understanding. *Technical Report 151*, SRI International, 333 Ravenswood Ave, Menlo Park, Ca. 94025.

Joshi A. K., and Kuhn S. (1979). Centered logic: The role of entity centered sentence representation in natural language inferencing. *In: International Joint Conference on Artificial Intelligence*.

Joshi A. K., and Weinstein S. (1981). Control of inference: Role of some aspects of discourse structure – centering", *In: International Joint Conference on Artificial Intelligence*, pp. 385-387.

Kennedy, C., Boguraev, B. (1996) Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. In: *16th International Conference on Computational Linguistics COLING'96*, Copenhagen, Denmark, pp. 113–118.

Lappin S., and Leass H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics* 20 (4), pp. 535-561.

McCallum A., and Wellner. B. (2003). Toward conditional models of identity uncertainty with application to proper noun coreference. In *Proceedings of the IJCAI Workshop on Information Integration on the Web*, pp. 79–84.

McCarthy, J. F. and Lehnert, W. G. (1995). Using decision trees for coreference resolution. In C. Mellish (Ed.), *Fourteenth International Conference on Artificial Intelligence*, pp. 1050-1055

Mitkov R. (1998). Robust pronoun resolution with limited knowledge. *In: 17th International Conference on Computational Linguistics* (COLING' 98/ACL'98), Montreal, Canada, pp. 869-875.

Mitkov, R. (1997). "Factors in anaphora resolution: they are not the only things that matter. A case study based on two different approaches". *In Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, Madrid, Spain.

Murthy K.N., Sobha L, Muthukumari B. (2007). Pronominal Resolution in Tamil Using Machine Learning Approach. *The First Workshop on Anaphora Resolution (WAR I)*, Ed Christer Johansson, Cambridge Scholars Publishing, 15 Angerton Gardens, Newcastle, NE5 2JA, UK, pp.39-50.

Ng V., and Cardie C. (2002). Improving machine learning approaches to coreference resolution. *In. 40th Annual Meeting of the Association for Computational Linguistics*, pp. 104-111.

Prasad R., and Strube,M.,(2000). Discourse Salience and Pronoun Resolution in Hindi, *Penn Working Papers in Linguistics*, Vol 6.3, pp. 189-208.

Ram, R.V.S. and Sobha Lalitha Devi. (2013)."Pronominal Resolution in Tamil Using Tree CRFs", *In Proceedings of 6th Language and Technology Conference, Human Language Technologies as a challenge for Computer Science and Linguistics* - 2013, Poznan, Poland

Recasens M., M`arquez L., Sapena E., Mart´I M.A., Taul´e M., Hoste V., Poesio M., Versley Y. (2010). SemEval-2010 Task 1: Coreference Resolution in Multiple Languages. In Proceedings of the *5th International Workshop on Semantic Evaluation, ACL 2010*, Uppsala, Sweden, .pages 1–8.

Recasens M., Hovy E. (2009). A Deeper Look into Features for Coreference Resolution. Lalitha Devi, S., Branco, A. and Mitkov, R. (eds.), *Anaphora Processing and Applications (DAARC 2009)*, LNAI 5847, Springer-Verlag Berlin Heidelberg, pp 535-561.

Rich, E. and LuperFoy S., (1988) An architecture for anaphora resolution. *In: Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas.

Senapati A., Garain U. (2013). GuiTAR-based Pronominal Anaphora Resolution in Bengal. *In: Proceedings of 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria pp 126–130.

Sikdar U.K, Ekbal A., Saha S., Uryupina O., Poesio M. (2013). Adapting a State-of-the-art Anaphora Resolution System for Resource-poor Language. *In proceedings of International Joint Conference on Natural Language Processing*, Nagoya, Japan pp 815–821.

Sobha L. and Patnaik B. N. (2000). Vasisth: An Anaphora Resolution System for Indian Languages. *In Proceedings of International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications*, Monastir, Tunisia.

Sobha L. and Patnaik,B.N. (2002). Vasisth: An anaphora resolution system for Malayalam and Hindi. *In Proceedings of Symposium on Translation Support Systems.*

Sobha L. (2007). Resolution of Pronominals in Tamil. Computing Theory and Application, *The IEEE Computer Society Press*, Los Alamitos, CA, pp. 475-79.

Sobha L., Pralayankar P. (2008). Algorithm for Anaphor Resolution in Sanskrit. *In Proceedings of 2nd Sanskrit Computational Linguistics Symposium*, Brown University, USA, 2008.

Sobha, Lalitha Devi., Vijay Sundar Ram and Pattabhi RK Rao. (2011). Resolution of Pronominal Anaphors using Linear and Tree CRFs. *In. 8th DAARC*, Faro, Portugal, 2011.

Sobha L., Sivaji Bandyopadhyay, Vijay Sundar Ram R., and Akilandeswari A. (2011). NLP Tool Contest @ICON2011 on Anaphora Resolution in Indian Languages. *In: Proceedings of ICON 2011*.

Soon W. H., Ng, and Lim D. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics* 27 (4), pp.521-544.

Strube, M. and Hahn U., (1999). Functional centering: Grounding referential coherence in information structure. *Computational Linguistics*, 25(3) pp 309–344

Taku Kudo. 2005. CRF++, an open source toolkit for CRF, *http://crfpp.sourceforge.net* .

Uppalapu. B., and Sharma, D.M. (2009). Pronoun Resolution For Hindi. *In: Proceedings of 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 09)*, pp. 123-134.

Wick M., Singh S., and McCallum A. (2012). A Discriminative Hierarchical Model for Fast Coreference At Large Scale. *In: Proceedings of ACL 2012*.