A Hybrid Approach for Anaphora Resolution in Hindi

Praveen Dakwale
LTRC, IIIT-Hyderabad
India
dakwale.praveen@gmail.com

Vandan Mujadia
CSPIT, Charusat
Gujarat, India
vmujadia@gmail.com

Dipti M Sharma
LTRC, IIIT-Hyderabad
India
diptims@gmail.com

Abstract

In this paper we present a hybrid approach to resolve Entity-pronoun references in Hindi. While most of the existing approaches, syntactic as well as data-driven, use phrase-structure syntax for anaphora resolution, we explore use of dependency structures as a source of syntactic information. In our approach, dependency structures are used by a rule-based module to resolve simple anaphoric references, while a decision tree classifier is used to resolve more ambiguous instances, using grammatical and semantic features. Our results show that, use of dependency structures provides syntactic knowledge which helps to resolve some specific types of references. Semantic information such as animacy and Named Entity categories further helps to improve the resolution accuracy.

1 Introduction

In various approaches on anaphora resolution syntax has been used as an important feature. Some well-known syntax based approaches include Hobbs algorithm (Hobbs, 1986) and the Centering approach (Brennan et al., 1987). Various rule based and data driven approaches have been proposed which use syntactic information as an important feature.

Most of the earlier works have used phrase-structure parse as a source of syntactic information. However, dependency structures are more suitable representations for relatively free word order languages such as Hindi (Bharati et al., 1995; Melčuk, 1988) and hence research in many such languages has focused on development of dependency based resources resulting in better availability of dependency data for such languages. In this paper, we explore the possibility of using dependency structure for anaphora resolution in Hindi. However, we do not intend either to propose dependency as an alternative to phrase structure or to compare the usability of the two frameworks.

(Prasad and Strube, 2000) is one of the most important approach for anaphora resolution in Hindi. They applied a discourse salience ranking to two pronoun resolution algorithms, the BFP and the S-List algorithm. (Dakwale and Sharma, 2011) reported the best performance for Hindi in Anaphora Resolution tool contest in Indian languages(ICON-2011). They propose a hybrid approach with limited linguistic knowledge such as NER categories and verb similarity.

Two earlier approaches explore the use of dependency relations for anaphora resolution. For Hindi, (Uppalapu and Sharma, 2009) extends the S-List algorithm by using two different lists in place of a single list. For English, (Björkelund and Kuhn, 2012) explores the possibility of using dependency relations as a feature for co-reference resolution in a learning based approach. Both the approaches are limited in their exploration of dependency for anaphora resolution as they only use dependency relations either as a salience for ranking the candidate referents or as an additional feature in a learning based approach. We discuss how resolution of different types of Entity-pronoun references in Hindi can benefit from dependency relations and semantic information. We present a hybrid approach to resolve Entity-pronoun references in Hindi which combines a rule-based system that uses dependency structures and relations and further improvement is achieved with semantic information such as animacy.

2 Data set and Grammatical Framework

In this work we use the data from the ‘Hindi/Urdu Dependency Treebank’ (Bhatt et al., 2009). It is a rich corpus with various linguistic information. The dependency annotation in this treebank is based on the Computational Paninian Grammar

International Joint Conference on Natural Language Processing, pages 977–981,
Nagoya, Japan, 14-18 October 2013.

(CPG-henceforth) framework, as is explained in (Begum et al., 2008) and (Bharati et al., 1995). This framework is based on the notion of '*karaka*' which are syntactio-semantic relations representing the participant elements in the action specified by the verb and it emphasizes the role of case endings or markers such as post-positions and verbal inflections. [1]. Table 1 shows some of the relevant relations and their rough correspondence to the traditional grammatical relations in English.

Label	CPG relation	Grammatical/thematic equivalent
k1	*karta*	Subject
k2	*karma*	Object
k4	*sampradan*	Experiencer/reciever
k7p(or k2p)	*adhikaran*	Location
r6	*sambandh*	Genitive

Table 1: CPG relations and equivalents

We use a part of the treebank which is also annotated with animacy information for Noun phrases as described in (Jena et al., 2013). Also, we used NE-Recognizer for Hindi to get the Named entity categories. The treebank has been annotated with anaphora links for all the pronouns as per the scheme described in (Dakwale et al., 2012). The size of the data that we use for our experiments is 325 documents, containing 4970 pronouns out of which 3233 pronouns are annotated as entity pronouns

3 System Description

The hybrid approach in our system is different from other hybrid approaches, in the way that instead of a rule based filtering followed by instance classification, our system includes a rule-based resolution module followed by a decision tree classifier for the remaining unresolved pronouns.

Anaphoric reference type can be classified into abstract (event) references where an anaphora refers to an event or a proposition and concrete (entity) references where it refers to a concrete entity like noun phrase (person,place etc), quantifiers etc. In this work, we focus on resolving only entity pronouns, hence the mention detection or anaphoricity determination step is not required for our system. Certain pronominal forms based on their different syntactic behaviour, can be resolved quite successfully with some specific rules using the dependency information. Therefore, we categorize pronominal forms in four types: Reflexive, Locative, Relative and Personal pronouns. The pronouns which are identified as concrete in the data are passed to the rule-based resolution module in which different rules depending on the category of the pronoun are applied to identify the correct referent. If none of the possible rules apply to a pronoun, it is passed to the classifier which uses a learning algorithm to identify the referent.

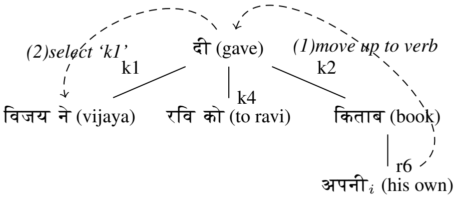

Figure 1

3.1 Rule based resolution module

The rule based module attempts to resolve the pronoun, using the dependency relations and other information, based on the category of the pronoun, which is decided using an exhaustive list of pronoun categories. We describe below some of the important rules used for different pronoun categories.

3.1.1 Reflexives

In Hindi *Possessive reflexives* are the most frequent reflexives which are only used in possession relation within the same clause and are different from third person possessive pronouns. Unlike English reflexives, they are not inflected with the gender and number of the possessor, but that of the possession. They include [अपना (*apana*), अपनी (*apanii*), अपने (*apane*)] (own). There are *Non-possessive reflexives* which can be used in any participant position, but mostly used in object position. They include [अपने आप(apane-aap), स्वयम्(swayam), खुद(khud)] representing 'oneself'. As it can be well derived from the binding theory, the referent of the reflexive pronoun is the accessible subject in its own governing category. Also, the 'k1' relation of CPG-based framework roughly corresponds to 'SUBJECT' of the traditional framework, thus the referent of the reflexive pronoun in most cases is the 'k1' of the same clause, i.e. that child node of the root verb of the clause, which has a dependency relation 'k1'.

(1) विजय$_i$ ने रवि को अपनी$_i$ किताब दी
vijay.ERG ravi.DAT his.POSS.REF book gave

'vijay$_i$ gave (his own)$_i$ (POSS.REF) book to ravi.'

[1] The detailed description of these relations is given in Hindi Dependency tagset (http://ltrc.iiit.ac.in/MachineTrans/research/tb/dep-tagset.pdf)

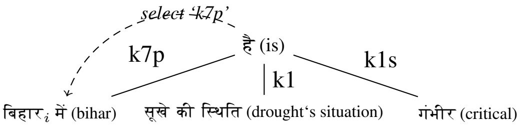

Figure 2

Figure (1) shows the dependency structure of example (1). The root verb of the clause containing possessive reflexive अपनी(*his*) is दी(*gave*) which has a descendant node विजय(*vijay*) with a dependency relation ‘k1’ with the verb. Thus it should be selected as the referent of the pronoun.

3.1.2 Place pronouns

Locative pronouns refer to location or places. They include वहां (‘there’) and यहां (‘here’). In CPG-based framework (as discussed in section 2), separate labels are used to represent the locative case, thus it can help in identifying the referents of these pronouns. To resolve place pronouns, we use dependency relations and Named entity Categories. Thus, place pronoun can be resolved by selecting the noun phrase nearest to the pronoun which has ‘LOCATION’ as NER-Category or the nearest NP with the dependency label ‘k7p’ or ‘k2p’. For ex :

(2) [$_{\mathrm{NER=LOC}}$ बिहार$_i$ में] सूखे की स्थिति गंभीर है। आज
bihar.LOC drought‘s situation critical is. today
प्रधानमंत्री ने वहां$_i$ का दौरा किया।
prime minister there visited

‘Situation of drought is critical in bihar$_i$. Today Prime minister visited there$_i$.’

Figure (2) shows the dependency structure of example(2). Noun phrase with NER category as ‘LOCATION’ nearest to the pronoun वहां (*there*) is (*bihaara*), thus it can be selected as the referent. In absence of NE category, dependency relations can be used to identify the referent.

3.1.3 Relative pronouns

In Hindi, relative pronouns include जो (which) and its case forms such as जिसे (to which), जिससे (from which) etc. In the CPG-based framework, relative clauses are marked with a relation ‘nmod-relc’, i.e. the relative clause is attached below that noun phrase which is relativized by the clause and the relation is labeled as ‘nmod-relc’. Thus, the referent of the relative pronoun should be selected as the noun-phrase to which the clause containing relative pronoun is attached. Consider following example

(3) बदमाशों से दस बैग$_i$ बरामद हुए जिनमें$_i$ वे चोरी का
From thugs ten bags restored in which they looted
सामान ले जाते थे
items used to carry

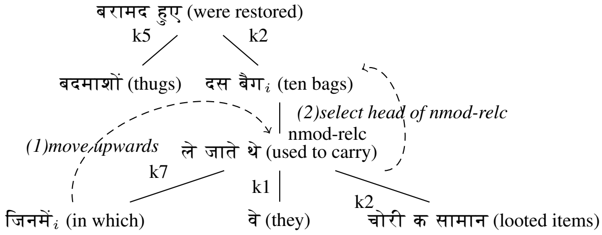
बरामद हुए (were restored)
k5
k2
बदमाशों (thugs)
दस बैग$_i$ (ten bags)
(2)select head of nmod-relc
nmod-relc
(1)move upwards
ले जाते थे (used to carry)
k7
k1
k2
जिनमें$_i$ (in which)
वे (they)
चोरी क सामान (looted items)

Figure 3

‘Ten bags$_i$ were restored from the thugs in which$_i$ they used to carry the looted items.’

Figure (3) shows the dependency structure of example (3), in which the relative pronoun is (*‘which’*) and the head of the relative clause is the verb ‘ले जाते थे’ (*used to carry*) which in turn is attached below the NP node (*‘ten bags’*) with a relation ‘nmod-relc’, which is selected as the referent of the relative pronoun.

3.1.4 Personal pronouns

All personal pronouns in Hindi are marked for number, respect and case. We consider resolution of first and second person pronouns seperate from third person pronouns.

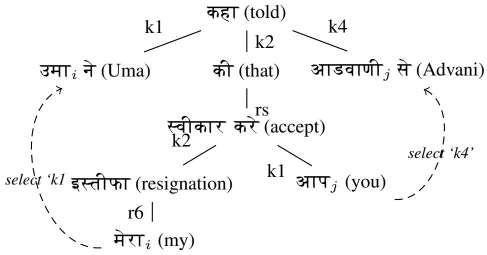

Figure 4

In the news corpus data, first and second person pronouns mostly occur in the narrative or attributional clauses (those subordinate clauses whose main clause has an attribution root verb such as बोल(*to tell*), कह(*to say*), बता(*to tell*) etc). If the first person pronoun is a part of attributional clause then its reference is the speaker of that clause. It is almost always ‘k1’ of the main clause. Similarly the referent of a second person pronoun in an attributional clause, is mostly the ‘k4’ or experiencer of the main clause. For ex :

(4) उमा$_i$ ने आडवाणी$_j$ से कहा की आप$_j$
Umaa.ERG advaani.DAT said that you.HONORIFIC.ACC
मेरा$_i$ इस्तीफा स्वीकार करे
my resignation accept

‘Umaa$_i$ said to advaani$_j$ that you$_j$ accept my$_i$ resignation.’

Figure (4) shows the dependency structure of example (4), in which (*you*) is second person pronoun in the attributional clause rooted at (*accept*), hence its referent is selected as the ‘k4’ of the main clause i.e. (*advani*). Similarly for the first person pronoun (*my*), referent is selected as ‘k1’ of the main clause i.e. (*Umaa*).

References of third person pronouns mostly are inter-clausal or inter-sentential. For third person anaphora, we adopt re-ordering of the candidate elements based on the salience of dependency relations, similar to (Uppalapu and Sharma, 2009). However, with two modifications : First, they consider the salience ordering (*k1 >k2 >k3 >k4 >others*) to rank the candidate elements, similar to ordering of the grammatical relation (*subject* >direct object *>indirect object >adjunct*) as in (Prasad and Strube, 2000). We adopt a slightly modified ordering of the relations (*k1 >k2 >r6 >k4 >k3 >others*) based on the relative frequency of the dependency relations for animate entities. Second, we also use animacy along with number to prune the candidate NP list. For ex :

(5) [$_{\mathrm{k1,h}}$ उमा$_i$ ने] [$_{\mathrm{k7}}$ पहले] [$_{\mathrm{k4,h}}$ शिवराज$_j$ को] [$_{\mathrm{k2,rest}}$ पत्र]
Uma.ACC first shivraaj.ACC letter
लिखा। फिर [$_{\mathrm{k1,h}}$ उन्होंने$_i$] [$_{\mathrm{k4,h}}$ उन्हें$_j$] [$_{\mathrm{k3,rest}}$ कोर्ट से]
wrote. later she.HON.ACC him.HON from court
[$_{\mathrm{k2,rest}}$ नोटिस] भिजवाया
notice sent.CAUSATIVE

‘First uma$_i$ wrote a letter to shivraaj$_j$. Later she$_i$ sent a court notice to him$_j$’

In the above example there are two pronouns in the second sentence : first is उन्होंने(*she*) (gender neutral). Salience based ordering of the possible referents for this pronouns is : [(*umaa*), पत्र (*letter*),(*shivaraj*)]. Since the top element i.e. (*umaa*) agrees with the pronoun in number and animacy, it is selected as the referent for (*she.ACC*). Thus the ordered list of candidates becomes : [पत्र (*letter*), (*shivaraj*)]. The second pronoun is (*him*) (gender neutral), but the top element in the list (*letter*) doesn’t agree with pronoun either in number or animacy. However, the next element in the list (*shivaraj*) agrees with the pronoun for both features, hence it is selected as the referent of the pronoun. If a pronoun could not be resolved within the two sentence, it is passed for learning based resolution.

3.2 Classifier module

We use the approach of (Soon et al., 2001) for classification. For training, a positive instance is created by pairing each anaphora and its actual antecedent, and negative instances are created by pairing the anaphora with multiple preceding non-antecedent Noun phrases. For testing, unlabeled instances are created by pairing the anaphora with all the Noun-phrases in upto 3 previous sentences. Testing instances are classified as positive or negative based on the model learned in the training phase. Positively labeled instances are then re-ranked based on the decision-tree-confidence-factor as described in (Witten and Frank, 2005). The NP-candidate corresponding to the highest ranked instance is proposed as the referent of the pronoun.

3.2.1 Features

Following features are used for classification:

- Number : singular, plural, honorific
- Named Entity categories: ‘Person’, ‘Organization’, ‘Location’, ‘Number’
- Distance feature: #NP chunks and #sentences between the pronoun and the candidate NP.
- Animacy : ‘human’, ‘animate’, ‘rest’.

4 Evaluation

We divide the treebank data approximately into ratio 2:1 for training and testing respectively. The training data contains 2162 entity pronouns and the test data contains 1071 entity pronouns.

4.1 Results

Table (2) shows the accuracies for different types of pronouns resolved by the rule-based module.

	Total pronouns	Correct Resolved	Accuracy
Reflexive Pronoun	156	129	.82
Relative Pronouns	80	68	.85
Locative Pronouns	48	37	.77
1st and 2nd person Pronouns	81	76	.93
Third person Pronouns	706	343	.48
Ovearll (Rule based system)	**1071**	**653**	**.60**

Table 2: Accuracy of the rule-based system

Results in Table 2 show that performance of the rule based system is quite high for all types of pronouns except for third person personal pronouns. This motivates us to use a learning based approach for the pronouns which remain unresolved in the first module. Table (3) shows the overall performance of the hybrid system achieved over the rule-based system by using different sets of features. The best performance (**0.70**) is achieved with a combination of all the features.

	Total	Correct	Accuracy
Rule based system(RB)	1071	653	.60
RB+Distance	1071	696	.64
RB+Distance+Agreement	1071	713	.66
RB+Distance+Animacy	1071	731	.68
RB+Dist+Animacy+Agreement	**1071**	**753**	**.70**

Table 3: Accuracy of the hybrid system

We provide a tentative comparison of our approach with two earlier systems : (Dakwale and Sharma, 2011) and (Uppalapu and Sharma, 2009).

Though an exact comparison is not possible due to unavailability of the data used in those systems.

	Total	Correct	Accuracy
Uppalapu-S	142	123	.86
Uppalapu-L	100	64	.64
(Dakwale and Sharma, 2011)	258	134	.52
Our system	**1071**	**753**	**.70**

Table 4: Resolution results of the three systems, Uppalapu-S and Uppalapu-L are the results of (Uppalapu and Sharma, 2009) for short and long story data respectively

4.2 Discussion and Error analysis

Table (4) shows that our system has achieved noticeable improvement over (Dakwale and Sharma, 2011), which is a knowledge poor approach using limited information. However, our system uses treebank data with information such as dependency and animacy.

(Uppalapu and Sharma, 2009) presents results for two sets of data, i.e. long and short stories. The overall accuracy of our approach is better than the accuracy for their long story data, although it is lower than theirs for their short story data. We have presented our results on treebank data which contains news articles from various domains with average size of 20 sentences, above results show that our approach performs consistently better even for longer texts and domain independent data. Also, the performance of the system for third person pronouns is relatively lower than that of other types of pronouns. Table 5 shows a breakup of the distribution of third person pronouns into two forms: Proximal and Distal, and their accuracies. The accuracy for resolution of proximal pronouns is exceptionally low than that of distal pronouns which can be attributed to the ambiguity in the resolution of distal pronouns which can refer to animate as well as inanimate objects

	Total	Correct	Accuracy
Proximal	132	43	.32
Distal	574	394	.68
Total Third person	706	437	.61

Table 5: Seperate results for proximal and distal third person

5 Conclusion and Future Work

The rule based system achieved a substantial accuracy of 60% which implies that dependency relations can help achieve an acceptable resolution performance for Hindi, and the use of decision tree classifier demonstrated a substantial improvement of 10% over the rule based system‘s accuracy. This shows that semantic features like animacy and Named entity categories provide important linguistic information for anaphora resolution.

In the current work, we have focused only on the resolution of entity pronoun references. In future we aim at the identification of reference type and resolution of event anaphora. We also aim to conduct experiments with dependency structures for anaphora resolution in other Indian languages such as Telugu, Bengali etc.

References

Rafiya Begum, Samar Husain, Arun Dhwaj, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. 2008. Dependency annotation scheme for indian languages. In *Proceedings of IJCNLP*.

Akshar Bharati, Vineet Chaitanya, Rajeev Sangal, and KV Ramakrishnamacharyulu. 1995. *Natural language processing: a Paninian perspective*. Prentice-Hall of India.

Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. A multi-representational and multi-layered treebank for hindi/urdu. In *Proceedings of the LAWIII*. ACL.

Anders Björkelund and Jonas Kuhn. 2012. Phrase structures and dependencies for end-to-end coreference resolution. In *Proceedings of COLING 2012*.

Susan E Brennan, Marilyn W Friedman, and Carl J Pollard. 1987. A centering approach to pronouns. In *Proceedings of 25th ACL*.

Praveen Dakwale and Himanshu Sharma. 2011. Anaphora resolution in indian languages using hybrid approaches. In *NLP tool contest, ICON 2011*.

Praveen Dakwale, Himanshu Sharma, and Dipti M Sharma. 2012. Anaphora annotation in hindi dependency treebank. In *Proceedings of PACLIC-26*.

Jerry Hobbs. 1986. Resolving pronoun references. In *Readings in natural language processing*. Morgan Kaufmann Publishers Inc.

Itisree Jena, Riyaz Ahmad Bhat, and Sambhav Jain. 2013. Animacy annotation in hindi treebank. In *Proceedings of the LAWVII*. ACL.

Igor Aleksandrovič Melčuk. 1988. *Dependency syntax: theory and practice*. State University of New York Press.

Rashmi Prasad and Michael Strube. 2000. Discourse salience and pronoun resolution in hindi. *U. Penn Working Papers in Linguistics*.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27.

Bhargav Uppalapu and Dipti Misra Sharma. 2009. Pronoun resolution for hindi. In *DAARC-7*.

Ian H Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.