Resolution of Pronominal Anaphora for Telugu Dialogues

Hemanth Reddy Jonnalagadda
IIIT Hyderabad
hemanth.reddy
@research.iiit.ac.in

Radhika Mamidi
IIIT Hyderabad
radhika.mamidi
@iiit.ac.in

Abstract

The challenge of anaphora resolution has been taken up from long time. However, most of the work did not include for dialogues. In this paper we discuss the types of pronouns and anaphora in Telugu language and make an attempt to build a rule based pronominal anaphora resolution algorithm for human to human conversations. The model mainly consists of two parts, creating a knowledge base with a set of pronouns along with its morphological information and designing an algorithm which uses this knowledge base to give an output. In this process we have worked on normal pronominal anaphora and suggested a set of rules applicable for Telugu dialogues. However, since there was no corpus for the Telugu language, we built a corpus and tested the algorithm on it. Results show that the suggested algorithm produced an output with an accuracy of 61.1%.

1 Introduction

Anaphora is the linguistic unit which refers to an entity that is previously mentioned in the same discourse. The word or phrase which it refers is called antecedent. The process of identifying the antecedent of an anaphora is called anaphora resolution. We find the importance of anaphora resolution in various NLP applications such as dialogue systems, text summarization, machine translation and so on. The pronouns in Telugu language carry much more information about the nouns than pronouns in English. Telugu is a morphologically rich and free word order language. The morphological richness of the language helps in building various NLP applications using simple parsing tools. Three types of anaphora are identified in Telugu language, they are normal pronominal anaphora, discourse deictic anaphora and one - anaphora which are discussed in detail in section 3. This paper concentrates mainly on normal pronominal anaphora. Many approaches have been proposed for anaphora resolution in discourse and in dialogues for different languages. This is the first attempt at anaphora resolution in Telugu dialogues.

2 Related work

Research has been going extensively in this area from past few decades. Many rule based, machine learning based and knowledge based systems have come up. One of the earliest work has been done by Hobbs (1976) on pronominal coreference. Hirst (1981) has compared 5 different approaches and presented their strengths, weaknesses in resolving the anaphora. Lappin and Leass (1994) came up with salience feature based approach. Mitkov (1994a) came up with integrated anaphora resolution approach which relies on set of preferences and constraints. Robust knowledge poor pronoun resolution was developed by Mitkov (1998) which makes use of part-of-speech tagger and simple noun phrase rules. Strube and Eckert (1998) made an attempt to resolve discourse diectic anaphora in dialogues. et al., (2004) built an anaphora resolution method for multi person dialogues where they brought in semantic, syntactic and knowledge based techniques which were used in anaphora resolution.

Coming to the Indian languages, work has been done mostly on Hindi, Bengali, Tamil and Malayalam. Shobha and Patnaik (2002) came up with a rule based approach VASISTH which currently handles anaphora in Malayalam and Hindi languages. Shobha.L (2007) resolves pronominals in Tamil, in which they have adopted a probabilistic model. Praveen et al., (2013) built a hybrid approach for anaphora resolution in Hindi using dependency parser and decision tree classifier. A generic anaphora resolution engine using machine learning was developed for Indian languages by Shobha et al., (2014), they have tested the system on Bengali, Hindi and Tamil. In Telugu Chandramohan and Sadanandam (2014) implemented a rule based pronominal anaphora resolution system for discourse. The algorithm was tested on a limited set of data and an accuracy of 60.75% was obtained.

3 Anaphora in the context of Telugu language

Telugu belongs to the Dravidian family of Indian languages. It is a free word order language. Two or more words undergoing some modifications to form a single compound word is called ***sandhi***. It is very common in agglutinative languages. In Telugu an entire sentence can be expressed as a single word which can be seen in the following example where three words are combined to form a single word.

Example 1:

nuvvekkaDunnAvu - nuvvu + ekkaDa + unnAvu

(Where are you?) you where present

These words cannot be analyzed by the NLP applications. To handle these cases there is a need for sandhi splitter which splits the word as shown above.

In Telugu, verbs are inflected with gender, number and person which give information about the subject,

Example 2:

The verb "vellu" can take all the below forms in Telugu. When subject is

male, sg, 3rd person - veltunnADu
female, sg, 3rd person - veltundi
neutral, pl, 3rd person - veltunnAru
neutral, sg, 1st person - veltunnAnu

3.1 Pronouns in Telugu

There are a wide variety of pronouns in Telugu. A pronoun can take different forms depending on gender, number, person, case and also on social relationships such as informal, formal, impolite, very polite etc.

This is illustrated below where all the pronouns refer to a male person relatively distant from the speaker.

vADu - 3rd person, sg, male, impolite
atanu - 3rd person, sg, male, neutral
Ayana - 3rd person, sg, male, polite
vAru - 3rd person, sg, male, very polite

There are 4 different types of pronouns in Telugu namely personal pronouns, demonstrative pronouns, reflexive pronouns and interrogative pronouns. The proposed algorithm works on all the above types of pronouns, excluding interrogative pronouns. In this paper we have classified these pronouns into two different categories and formed rules based on this.

Personal pronouns: Pronouns which map to human characters

Example 3:

"atanu" *(he)*, "Ame" *(she)*, "vADu" *(he)*, "tamaru" *(you)*, "adi" *(she*), "nuvvu" *(you)*, "nEnu" *(me)* etc.

Non-personal pronouns: These pronouns map to the objects or animals

Example 4:

"adi" *(it)*, "avi" *(they)*, "vATiki" *(them)* etc.

3.2 Types of Anaphoras in Telugu Dialogues

Normal pronominal Anaphora: Where pronoun is referring to single word which is previously introduced in the context.

Example 5:

ramaNa: rAjA elA unnAvu? ninnu
(Raja how present you)
cUsi cAlA rojulayiMdi.
(seen many days back)

(Ramana: How are you? It's been long time since I met you)

rAjA: **nEnu** bAgunnAnu, nuvvela unnAvu?
(me good you_how present)

(Raja: I am fine. How are you?)

Discourse Diectic Anaphora: In this case, there is no single antecedent possible. The pronoun will be referring to multiple words in the previous sentence. This is illustrated in example 7 where the pronoun "adi" *(that)* is referring to the entire dialogue.

Example 6:

rAmu: **rAju cAla mancivADu.**
(Raju very good person.)
(Ramu: Raju is a very good person)
ravi: **adi** nIku ela telusu?
(that you how know)
(Ravi: How do you know that?)

One - Anaphora: It is the anaphoric noun phrase headed by the word one (okaTi in Telugu).

Example 7:

rAju: nEnu kotta **baMDi** koMdAmani
(I new bike buying)
anukuMTunnAnu.
(*thinking.)*
(Raju: I am thinking of buying a new bike)
ravi: nAku **okaTi** kAvali.
(I one need.)
(Ravi: I too need one.)

4 Anaphora Resolution

Our approach for anaphora resolution can be explained in three steps

- Parsing the dialogues
- Creating the knowledge base and
- Identifying the antecedent.

4.1 Parsing

We parse the dialogues one after the other using morph analyzer. In this process we store the speaker of each dialogue. The output of the morph analyzer is given as input to the POS tagger. The final output at this stage is in Shakti standard format (SSF). It is the highly readable representation for storing language analysis, an English and Telugu examples in SSF are given below.

```
children        NNS     <fs af='child,n,m,p,3,0,,' >
                                |    | | | | |
                                |    | | | |  \
                               root  | | |pers |
                                     | | |    case
                              category | number
                                       |
                                     gender
```

atanu PRP <fs af='atanu,pn,m,sg,3,d,0,0' name="atanu">

Initially we have implemented the system using only morph analyzer where POS tag of most of the pronouns, nouns and their features were not identified.
Example 8:
vATini
<fsaf='vATini,unk,,,,,,'poslcat="NM"name="vATini">

As we can see in this example, the category of this pronoun is shown as unknown and gender, number, person and case features are also null values.
The use of POS tagger in addition to the morph analyzer helped in obtaining POS tags for the nouns and pronouns which could not be tagged by the morph analyzer.

4.2 Knowledge Base

We stored all the pronouns which are tagged by POS tagger and not detected by the morph analyzer in a separate database file. Along with it we have also stored the gender, number, person and case features of these pronouns in the knowledge base. While running the anaphora resolution algorithm, if a pronoun falls in the above category then the information of that particular pronoun is obtained from the database file.
We have also stored the list of reflexive pronouns in the knowledge base as the algorithm consists of rules applicable only for reflexive pronouns.

4.3 Algorithm

We have implemented a rule based pronoun resolution algorithm. These rules were formed based on various pronouns and features obtained from the parsers. We have tried to concentrate more on 3rd person pronouns because it was relatively easy to identify the antecedents of 1st and 2nd person pronouns compared to 3rd person in human conversations. The system works on the rules explained below,

4.3.1 Personal pronouns

As mentioned in sec 3.1 personal pronouns should be referring to the human characters, these pronouns are mapped to the proper nouns or common nouns which tell about human identity.
Example 9:
"tammuDu" *(younger brother)*, "anna" *(elder brother)*, "akka" *(elder sister)*, "amma" *(mother)*, "DAktar" *(doctor)*, "tIcAr" *(teacher)*, etc.

1st person pronouns

1st person pronouns are mapped to the current speaker, if the root word of the pronoun is "mana" *(we)* then it will map to both speaker and listeners. Listeners are the speakers of previous and next dialogues.

Example 10:
amala: **nA** pEru amala. nI pEreMTi?
(My name Amala. You name_what)
(Amala: My name is Amala. What's your name?)
sIta: nA pEru sIta.
(My name Seetha.)
(Seetha: My name is Seetha.)
Example 11:
srInivAs: nAku edayina kottagA
(I something new)
nercukovAlani uMdi.
(learn there.)
(Srinivas: I want to learn something new.)
gOpAl: kAni **manaku** vIlavvutuMdA?
(but us feasible.)
(Gopal: But is it feasible for us?)

As we can see, in example 10 "nA" is referring to "amala" and in example 11 "manaku" which is

a plural pronoun is referring both “srInivAs” and “gOpAl”

2nd Person pronouns

2nd person pronouns are mapped to the listener, if there is a previous dialogue then it will refer to the speaker of the previous dialogue. If there is no previous dialogue then it will be mapping to speaker of the next dialogue.

Example 12:
sIta: nA pEru sIta.
(my name Seetha.)
(Seetha: My name is Seetha.)
amala: **nuvvu** ekkaDa caduvutunnAvu?
(you where studying?)
(Amala: Where are you studying?)

Example 13:
amala: nA pEru amala. **nI**
(My name Amala you)
pEreMTi?
name_what?
(Amala: My name is Amala. What’s your name?)
sIta: nA pEru sIta.
(my name Seetha)
(Seetha: My name is Seetha.)
“nuvvu” in example 12 and “nI” in example 13 are referring to the speakers of previous and next dialogues respectively.

3rd Person pronouns

We have tried to identify the antecedents based on gender, number, person features and a few rules. The step by step procedure and the rules are as follows,

1) While analyzing the parsed text word by word. If a possible antecedent is detected then it is stored along with its POS tag, gender, number and person.

   Example 14:
   ravi: **rAmuki** Dabbu evaru iccAru?
   (to_Ramu money who gave)
   (Ravi: Who gave the money to Ramu?)
   rAju: gOpi **ataniki** Dabbu iccADu.
   (Gopi him money gave)
   (Raju: Gopi gave him the money)

   All the nouns that appear before pronoun can be possible antecedents. Possible antecedents for the pronoun “ataniki” in example 14 are shown below with its morphological features,

   rAmuki NNP <fs af='rAmuki,unk,,,,,,' poslcat="NM" name="rAmuki">
   Dabbu NN <fs af='Dabbu,n,,sg,,d,0,0'name="dabbu_2">
   gOpi NNP <fs af='gOpi,n,,sg,,d,0,0' name="gOpi">
2) If a pronoun is detected then we will be extracting its gender, number and person features from the morph analyzer or through the preprocessing stage mentioned above.
   ataniki PRP <fs af='atanu,pn,m,sg,3,,ki,ki' name="ataniki">
3) From the list of all possible antecedents, the antecedents which do not agree in gender, number and person features with the pronoun are removed from the list. Below constraints will be applied on the remaining antecedents.
   As we can see the noun “dabbu” agree in gender, number and person with the pronoun. Though the features of “rAmuki” and “gOpiki” are not identified by the morph analyzer, we have considered it as the possible antecedent.
4) Pronouns which come under the category of personal pronouns will be mapped to the proper nouns or the common nouns which tells about human identity. Some of these common nouns are given in example 9.
   As “Dabbu” cannot be referred by the personal pronouns, it is removed from the possible antecedent lists, we are left with “rAmuki”, “gOpiki”
5) Reflexive pronouns will be referring the antecedents within the sentence. Example: “tanu” *(he/she)*,”tAmu” (they).
   Example 15:
   ravi: rAmuki Dabbu evaru iccAru?
   (to_Ramu money who gave)
   (Ravi: Who gave the money to Ramu?)
   rAju: **gOpi** **tana** Dabbu ataniki
   (Gopi his money him)
   iccADu.
   (gave.)
   (Raju: Gopi gave his money to him.)
6) Whereas demonstrative pronouns will be referring to the antecedents outside the sentence. Example: “vADu” *(he)*, “Ame” *(she)*, etc.
   The referent in the Example 14 is a demonstrative pronoun, it should be referring the antecedent outside the sentence. So “gOpi” is removed from the list and we are left with only one possibility “rAmuki” which is the antecedent of the pronoun.

7) If more than one antecedent is left in the list, recently appeared one is preferred as its antecedent.

4.3.2 Non-personal Pronouns

We have applied the same procedure as mentioned above but the steps 4, 5, 6 will change because these pronouns will be mapping to the objects or animals. So, there is no possibility of reflexive pronouns.

Example 16:
gOpi: ravi nI **kukkalu** elA unnAyi?
(Ravi your dogs how present)
vATiki AhAram peTTAvA?
(them foodkept?)
(Gopi: How are your dogs? Did you feed them?)
ravi: **avi** bAgunnAyi. **vATiki** AhAraM
(they good them food)
peTTAnu.
(Kept.)
(Ravi: They are good. I fed them.)

5 Building the corpus

We have noted all the possible pronouns possible for Telugu and for each pronoun, we considered all the possible forms which it can take in different cases (Genitive, Nominative, Objective, and Dative).

The corpus consists of human to human conversations which has all the above pronouns excluding interrogative pronouns, as it is not part of our research currently.

The corpus consists of 108 human conversations, each conversation may contain around 2-8 dialogues. Total number of pronouns in the corpus are 509. About 40% of the corpus has been taken from online chat conversations and the remaining 60% of the conversations have been taken from telugu.webdunia.com and www.learningtelugu.org.

6 Results

Once we built the corpus, we have tagged all the pronouns in the corpus with its corresponding antecedents. This has been done manually. Corpus consists of all categories of pronouns mentioned in sec 3. The algorithm has been run on the entire corpus and we have compared the results with the tagged data.
We have analyzed the results based on the person of the referents. The accuracy of the system on different pronouns is shown in the below table,

	No. of pronouns	Pronouns unresolved	Pronouns wrongly resolved	Pronouns correctly resolved	Accuracy
1st person	138	11	14	113	81.88%
2nd Person	106	14	19	73	68.87%
3rd person	202	23	91	88	43.56%
Non personal pronouns	63	3	23	37	58.73%
Overall	509	51	147	311	61.10 %

7 Issues in Telugu

Corpus

Annotated text and annotated dialogues are not available for Telugu language. All the data is in printed books format and the text which is present in most of the Telugu websites is in image format which makes it difficult to extract.

Sandhi

This phenomenon makes it difficult for the parser to analyze the text. In such cases the algorithm is not able to identify the antecedent or the anaphora present in the sentence. The accuracy of the parsers will be less resulting in the low accuracy of the system which was dependent on these parsers.

Example 17:
ravi: ninna sIta inTiki vacindi.

(yesterday seetha home came)
(Ravi: Yesterday, Seetha came to my home.)
rAju: AmeMdukocindi?
Ame + eMduku + vacindi.
(she why came.)
(Raju: Why did she come?)

In example 17 Pronoun is a part of the compound word in such a case it becomes difficult to identify the anaphora/antecedent. Hence we require a good sandhi splitter for parsing the text.

Subject identification

Telugu is a free word order language. The morph analyzer which we are using is taking individual word as an input, so we are unable to get the case of the words and are unable to identify the subject. Identifying the subject of the sentence will improve the accuracy of 3rd person pronouns.

Example 18:

rAju: raviki sIta pustakaM icindi.
(to_Ravi Seetha book gave)
ataniki pariksa undanta.
(he exam present.)
(Raju: Seetha gave the book to Ravi. He has an exam.)

In the above example, "sIta" is the subject of the first sentence and its gender is feminine which can be identified from the verb "icindi". As the algorithm is unable to identify the subject, the pronoun "ataniki" (male) is mapped to "sIta" (female).

Parsers

Morph analyzer and POS tagger used are giving lot of errors. These errors are carried forward affecting the overall accuracy of the anaphora resolution algorithm.

8 Future Work

In this paper we have concentrated mainly on normal pronominal anaphora. We would like to concentrate on cataphora in our future work as Telugu is a free word order language the possibility of occurrence of cataphoric reference is very high.

Work has been going on in dialogue systems in Telugu for tourism domain. We would like to add this anaphora resolution system into dialogue system.

We would like to work on 3rd person pronouns as it has very less accuracy.

Adding a good sandhi splitter to the system will improve the accuracy of the system.

References

Krishnamurti, Bhadriraju; J.P.L.Gwynn (1985). A Grammar of Modern Telugu. New Delhi: Oxford University Press.

http://web.cs.ucdavis.edu/~vemuri/Grammar/9.%20pronouns-1.pdf

Lappin & Leass (1994) Sh. Lappin, H. Leass - An algorithm for pronominal anaphora resolution. Computational Linguistics, 20(4), pp. 535-561.

Miriam Eckert & Michael Strube (1998). Discourse Deictic Anaphora in Dialogues. Institute for Research in Cognitive Science. University of Pennsylvania.

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.

Hobbs,J. R. (1976). Pronoun Resolution, Research Report 76-1, Department of Computer Science, City College, New York

D. Chandra Mohan and M. Sadanandam (2014), Telugu Pronominal Anaphora Resolution. International Journal of Research and Applications

Daniel Jurafsky and James Martin (2009), Speech and Language Processing, Prentice-Hall.

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

Sobha.L and Vijay Sundar Ram (2014) - A Generic anaphora resolution engine for Indian languages. In the proceeding of COLING.

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

Praveen Dakwale, Vandan Mujadia, Dipti M. Sharma (2013). "A Hybrid Approach for Anaphora Resolution in Hindi." LTRC, IIIT-Hyderabad. India.

Prateek Jain, Manav R. Mittal, A Mukerjee and Achla M. Raina, 2004. "Anaphora Resolution in Multi-Person Dialogue", Strube, M. and Candy Sidner (ed.), Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue.

Akshar Bharathi, Rajeev Sangal and Dipti Mishra (2007) Shakti Standard Format.