ANAPHORA RESOLUTION IN TAMIL

RULE-BASED APPROACH

Sumanth Balaji 2018114002

Tanvi Kamble 2018114004

Abstract

The following document describes the strategies used to resolve anaphora in Tamil Language and brief summaries of the papers referred in order to come up with a solution.

Table of Contents

1 Introduction 1

2 Anaphora in Tamil 2

3 Hobbs Algorithm 4

4 Algorithm 6

5 Accuracy and Scope for future work 8

6 Challenges Faced 9

7 References 11

1. Introduction

This project aims to resolve anaphora used in the Tamil Language typically from Short stories using a rule-based approach. We have analysed manually the corpus that contains data from the stories to generate rules to identify the anaphoric personal pronouns, if any.

2. Anaphora in Tamil

What is Anaphora?

Anaphoric expressions in natural language text help in bringing cohesion to the text. The resolution of these anaphoric expressions is vital in developing information extraction and understanding systems.

Types of Anaphora

1. Repeated NPs: Same word used in different sentences refers to the same entity. Example: John is a nice boy. John is five years old. –¿ Here both the Johns refer to the same person.

2. Partially Repeated NPs: Only a part of the first occurrence is repeated. Example:Prime Minister Narendra Modi is going to visit Hyderabad. Modi will stay there for two days. Here Modi refers to Prime Minister Narendra Modi

3. Lexical Substitution: When a synonymous word repeats in the further occurrences.
   Example: Mary is a talented artist. The child is a prodigy. Here The child refers to Mary.

4. Pronominal: The most common form of anaphora. When a noun is replaced by a pronoun in the repeating occurrences.
   Example: John met Mary. She seemed happy. She is referring to Mary

5. One Substitution: John bought a red car. Bill bought the red one. Here one in the second sentence refers to the car from the first one!

6. Ellided: Ellipses occur and the repeated noun is skipped.
   Example: John has a book. Mary has too. This means that Mary also has the book that John has.

Tamil Language

Tamil belongs to south Dravidian family of languages. It has post-positions. It is

nominative-accusative language like the other Dravidian languages. The subject of Tamil sentence is mostly nominative. There are constructions with certain verbs that require dative subjects and possessive subjects. Tamil has PNG (person, number and gender) agreement. A pronoun must agree in number, gender and person with antecedent. There are many types of anaphora. The types are pronominal anaphora, possessive anaphora, reflexive anaphora, demonstrative anaphora, relative anaphora. The following are few examples of pronouns in Tamil

1. Personal Pronouns - Tamil
   (a) *I* am your friend [**nāṉ** uṅkaḷ naṇpaṉ]
   (b) *you* speak very fast [**Nīngal** vēkamāka pēsukiṟīrkal]
   (c) *He* has three dogs Avarukku mu nyka uaa [**avaritam** mooṉṟu nāykaḷ uḷḷaṉa]
   (d) *she* can speak German [**avalal** jerman pesamutiyum]
   (e) *we* will not come late [**nangal** thāmadhamk̄a vara māttom]
   (f) *they* bought milk and bread [**avarkal** pāl matrum rotti vānginārkal]
2. Object Pronouns
   (a) can you tell *me* your name? [neengal ungal peyarai **enakku** solla mutiyum?]
   (b) I will give *you* money [nan **unakku** panam kotuppen]
   (c) she wrote *him* a letter [aval **avanukku** oru katidham ezhudhinal]
   (d) they visited her yesterday [avarkal netru **avalai** sendru parththanar]
   (e) can she help *us*? [aval namakku **udhava** mutiyuma?]
   (f) he gave them food [avan **avarkalukku** unavu kotuththun]

3. Hobbs Algorithm

Hobbs Algorithm is a rule based algorithm to resolve anaphora of English Language mainly personal and reflexive pronouns using the parse trees. Following is the given algorithm (as given here):

1. Begin at NP
2. Go up tree to first NP or S. Call this X, and the path p.
3. Traverse all branches below X to the left of p. Propose as antecedent any NP that has a NP or S between it and X
4. If X is the highest S in the sentence, traverse the parse trees of the previous sentences in the order of recency. Traverse left-to-right, breadth first. When a NP is encountered, propose as antecedent. If not the highest node, go to step 5.
5. From node X, go up the tree to the first NP or S. Call it X, and the path p.
6. If X is an NP and the path to X did not pass through the nominal that X dominates, propose X as antecedent
7. Traverse all branches below X to the right of the path, in a left-to-right, breadth first manner. Propose any NP encountered as the antecedent
8. If X is an S node, traverse all branches of X to the right of the path but do not go below any NP or S encountered. Propose any NP as the antecedent.

WHY DO WE NEED TO MODIFY HOBBS?

Tamil is a very morphologically rich language where one verb can take multiple forms in different contexts.

The Parse trees of Tamil and English are quite different. As you can see, in Tamil there is very less evidence of searching in a hierarchical way as almost all the nouns are present at the same level. So the hobbs algorithm doesnt make sense

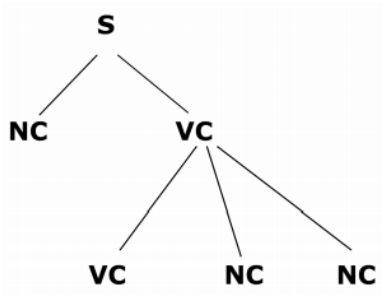

Fig. 2. Sample tree for English

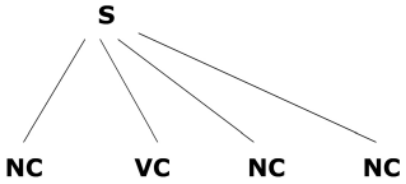

Fig. 3. Sample tree for Tamil

4. Algorithm

Pre-Processing

Creating Shallow Parse Trees from the Tamil data.

Due to lack of resources available for Tamil language, the code cannot parse the given text on its own. Instead the user is expected to feed already parsed data.

Since anaphora resolution mainly deals with NPs and VPs, shallow parsing also works.

For this project we use the ltrc shallow parser available here

Example:

Consider the sentence, "She went to his house" translated as *Avaḷ avaṉ viṭṭiṟku ceṉṟaḷ*

When run into the shallow parser we get:

```
1 ((        NP
    1.1 Avaḷ        PRP
))
2 ((        NP
    2.1 avaṉ        PRP
))
3 ((        NP
    3.1 viṭṭiṟku        NN
))
4 ((        VGNF
    4.1 ceṉṟaḷ        VM
))
```

Algorithm

1. Given a text, the algorithm first sorts out all the possible pronouns that can be resolved.

2. For each pronoun found, the code traverses in the reverse order searching for the potential antecedents in the preceding three sentences.

3. Once such a noun is found, two checks are made:
   (a) If the number of the noun and pronoun match (Singular/Plural)
   (b) If the gender of the noun and pronoun match (Male/Female) - The NLTK library is used for the same
4. If both of these features match, then claim that the noun is the antecedent to the pronoun

Categorization of pronouns:

```
nominal_labels = ["NN", "NNS", "NNP", "NNPS", "PRP"]

number_info = \
    {"NN": "singular",
     "NNP": "singular",
     "அவரை": "singular", #
     "அவளை": "singular", #
     "அவன்": "singular", #
     "அவர்": "singular", #
     "அவள்": "singular", #
     "அது": "singular", #
     "தன்னை": "singular", #
     "NNS": "plural",
     "NNPS": "plural",
     "அவர்கள்": "plural",#
     "அவர்களுக்கு": "plural", #
     "தங்களை": "plural", #
     "அவர்களே": "plural", #
     "PRP": None
     }

group_pronouns=["அவர்களே","தங்களை","அவர்களுக்கு","அவர்கள்"]
male_pronouns = ["அவர்","அவன்", "அவரை", "தன்னை"]
female_pronouns = ["அவள்", "அவளை", "தன்னை"]
neuter_pronouns = ["அது", "தன்னை"]
```

5. Accuracy and Scope for future work

The code is 72 percent Accurate.

The probable reasons are stated on the next page

Scope for future work

The code can be made more efficient by describing a method to resolve ambiguities with the help of SR Rules. Reflexive pronouns can also be taken into consideration. We can also modify the code to look for repetition as Anaphora.

Contribution

There was equal contribution made by both of us and is very difficult to draw a line to show individual work.

6. Challenges Faced

1. Being a morphologically rich and free language, the verbs in Tamil can take different forms depending on the context. There arent any resources online to obtain the SR Rules for Tamil and doing that manually is a practically impossible task! Therefore, we werent able to check the ambiguity by cross checking the SR rules of a verb with the ontology of the nouns.
   However here is an example of what we were planning to do:
   Given text:Kuḻantai taraiyil viḻuntatu. Avaṉ ōṭikkoṇṭiruntāṉ. which translates to- The child fell on the ground. He was running.
   Here we want to resolve the pronoun he Avaṉ.
   Ontology of child: PRSN, MML, FAUNA, ANIMT, N
   Ontology of ground: ABS, INANI, N
   The verb ōṭikkoṇṭiruntāṉ (running): Thematic roles: Agent SR rules for the agent : +Animate
   Since only child in animate we declare that the pronoun he was referring o the child without even the need for hobbs algorithm
   But different forms of Running in tamil: ōṭikkoṇṭiruntāḷ, ōṭikkoṇṭiruntātu, Iyankum etc. Therefore we cant use SR rules!

2. The shallow parser tool is not completely accurate but the best online parser available for Tamil language.It doesn't take into account the first word of the sentence This tool lets you input only one sentence ate a time so very time taking task if you want to parse an entire story

3. There aren't really any research papers that focus on anaphora resolution in Tamil that are rule based. Hence we did not have an idea as to how to make rules in dravidian languages

4. Tamil, being a linguistically rich language has a lot of variations in the same word without really making a difference semantically. This redundancy made it difficult for us to consider all the cases and the detecting which form of the word is used.

5. Lack of Resources: It was a difficult task to find data in Tamil which was translated in in English

7. References

1. ltrc.iiit.ac.in

2. *Enhancing Tamil WordNet with Subcategorization Information* Vijay Sundar Ram R and Sobha Lalitha Devi AU-KBC Research Centre, MIT Campus of Anna University, Chennai, India

3. *How to Handle Split Antecedents in Tamil?* Vijay Sundar Ram R. and Sobha Lalitha Devi AU-KBC Research Centre, MIT Campus of Anna University, Chennai, India

4. *Resolution for Pronouns in Tamil Using CRF* Akilandeswari, A and Sobha, Lalitha Devi AU-KBC Research Centre, MIT Campus of Anna University, Chennai, India

5. *A Generic Anaphora Resolution Engine for Indian Languages* Vijay Sundar Ram R. , Pattabhi RK Rao and Sobha Lalitha Devi AU-KBC Research Centre, MIT Campus of Anna University, Chennai, India

6. *Anaphora Resolution in Tamil Through Decision Tree Learning- Challenges and Tackles* Arul Deepa K*, Deisy C

7. http://learn101.org/tamil_grammar.php

8. Jurafsky Martin Speech and Language Processing Second Edition Draft 2007

9. http://www.cfilt.iitb.ac.in/indowordnet