Resolution for Pronouns in Tamil Using CRF

Akilandeswari, A and Sobha, Lalitha Devi
AU-KBC Research Center, MIT Campus of Anna University
Chennai,India
akila@au-kbc.org, sobha@au-kbc.org

ABSTRACT

The main goal of this paper is to develop an automatic anaphora resolution system for Tamil. Here we present the complete analysis of pronominals in Tamil discourse. We have analysed manually the corpus which contain Tamil pronominal aval, avan, atu and its suffixes. Using the analysis we come up with set of features to identify the anaphoric pronouns and its antecedents. We used the machine learning algorithm, Conditional Random Fields to this problem. The results are encouraging.

KEYWORDS : anaphora, pronoun, antecedent, pronominal, machine learning, conditional random fields.

Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), pages 103–112,
COLING 2012, Mumbai, December 2012.

1 Introduction

Anaphora resolution is an important task in natural language processing applications such as Information Retrieval, Information Extraction, Question Answering system, Text summarization etc. The process of finding the antecedent of an anaphor is anaphora resolution. Anaphora is the reference that point to the previous item. Antecedent is the entity to which the anaphor refers. The Tamil pronominals are '*avan*', '*aval*' and '*atu*' are the third person singular pronouns. 'atu' is third person singular neuter pronoun. The relation that is established between a pronoun and its antecedent helps to provide more information that can be extracted. Here we use conditional random fields (CRFs) to resolve this task, the results are encouraging.

2 Previous Work

Anaphora resolution is one of the difficult problems in the field of NLP. It is an area well researched for many languages in the last few decades. There are several approaches used in resolving pronominal such as rule based, knowledge based and machine learning.

One of the early works in pronominal resolution is by Hobb's naive approach, which relies on semantic information (Hobbs. J,1978). Carter with Wilkas' common sense inference theory came up with a system (Carter. D,1987). Carbonell and Brown's introduced an approach of combining the multiple knowledge system (Carbonell. J. G. & Brown. R .D, 1988). The initial approaches, where broadly classified as knowledge poor and rich approach. Syntax based approach by Hobb (naive approach), centering theory based approaches (Joshi, A. K. & Kuhn. S, 1979; Joshi, A. K. & Weinstein.S, 1981) and factor/indicator based approach such as Lappin and Leass' method of identifying the antecedent using a set of salience factors and weights associated to it. This approach requires deep syntactic analysis. Ruslan Mitkov introduced two approaches based on set of indicators, MOA (Mitkov's Original Approach) and MARS (Mitkov's Anaphora Resolution System) (Mitkov. R,1998). These indicators return a value based on certain aspects of the context in which the anaphor and the possible antecedent can occur. The return values range from -1 to 2. MOA does not make use of syntactic analysis, whereas MARS system makes use of shallow dependency analysis.

There are very few works done in anaphora resolution with respect to Indian Languages. Some of the works done are VASISTH a rule based system which works with shallow parsing and exploits the rich morphology in Indian languages for identifying the antecedent for anaphors (Sobha.L & Patnaik.B.N, 1999). (Sobha.L, 2007) used salience measure for resolving pronominals in Tamil. (Murthi.K.N, 2007) have looked into the anaphora resolution in Tamil, using Machine Learning technique: Linear Regression and compared it with salience factors. Dhar worked on "A method for pronominal anaphora resolution in Bengali (Dhar.A & Garain.U, 2008; Sobha.L & Pralayankar.P, 2008) worked on "Algorithm for Anaphor Resolution in Sanskrit". Resolving Pronominal Anaphora in Hindi Using Hobb's Algorithm was done by Kamalesh Dutta.

In ICON 2011, NLP tool contest on Anaphora Resolution for Indian Languages was held. The tool contest considered the languages such as Bengali, Hindi, Odiya, Marathi and Tamil. In each language different methods was approached by the participants.

3 Anaphora In Tamil

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. Generally anaphors have antecedents which we say as anaphoric. Anaphors which is not having antecedent in the text or does not refer any text before is non-anaphoric. We have taken Tamil pronominals He - *avan*, She - *aval*, It – *atu* and its suffixes. Let us classify the Tamil pronominals as reflexive pronoun, possessive pronoun and non-possessive pronoun. They are given below.

Reflexive Pronominals:

Himself	avane he+e(clitic marker)	avaNAka he+benefactive	
Herself	avaLe she+e(clitic marker)	avaLAka he+benefactive	
Itself	atuve It+e(clitic marker)	atAka It+benefactive	ate It+e(clitic marker)

The reflexive pronoun has two morphemes 'e' clitic marker and 'aaka' benefactive case marker.

Posessive Pronominals:

His	avan**utaiya** he+possessive	avan**atu** he+possessive	avan**in** he+possessive	
Her	aval**utaiya** she+possessive	aval**atu** she+possessive	aval**in** she+possessive	
Its	athanutaiya It+posseseive	athanatu It+posseseive	athanin It+posseseive	athan Its

Non-posessive Pronominals for (*avan*, *aval*):

avan**ukku** - He+dat, avaL**ukku** - she+dat. Similarly aval, avaL and atu is suffixed with case markers such as accusative,locative sociative,ablative, instrumental are considered to be non-possessive pronominal.

3.1 Examples for Tamil Pronominals

Example 1:

In this example, anaphora and antecedent are residing in the same sentence.

{***piiman***$_i$ cenru ciRaippattu} NF {pinpu krishnan ***avanai***$_i$ kaapaaRRinaar.}MCL
Bhiman went caught prisoned after Krishnan he+acc rescued
'Bhiman caught prison, after Krishnan rescued him.'

In the above example the antecedent is ***piiman***$_i$, which is subject and nominative. proper noun of previous clause. Even though krishnan is the nominative proper noun which is immediate to the anaphor ***avanai,*** krishnan cannot be the antecedent. A proper noun which is followed by a pronoun with or without case marker, cannot be the antecedent for the pronoun except in the case of example 8.

Example 2:

In this example, there are two sentences 2.0 and 2.1. The anaphora is in 2.0 sentence and the antecedent is in previous sentence 2.1.

2.1 **Intirajit**$_i$ RamarooTu yutatil tooRRuppooi}NF {ivitatil ampikaiyai vazhipaaTu
Indirajith Ramar+soc war+loc lost this place+loc ampikay+acc worshipped
ceytaan.}MCL
done+3msg
'Indirajith lost in the war when fought with Ramar, he worshipped goddess in this place'.

2.0 **{avan**$_i$ pinnar mooTcam aTaintaan.}MCL
He after wisdom reached+3sgm
'After he got wisdom'.

In the above example the antecedent is **Intirajit**$_i$, which is subject and nominative proper noun in the previous sentence to the anaphor. Even though *RamaroTu* is a proper noun with sociative case marker, it cannot be the antecedent. The proper noun **Intirajit**$_i$ with nominative case marker is the most probable antecedent. ***avan***$_i$.is the nominative anaphor.

Example 3:

3.3 {**Raaman**$_j$ aluvalakattiLiruntu viitirku vantu,}NF {**Siitaavai**$_i$ paartan.}MCL
Raman office+loc+ abl home+dat come sita+acc saw+3sgm
'Raman came from the office and saw Sita'.

3.2 {***avaL***$_i$ tanakuu uTampu cariyillai enRu connataal}CON {***avaLai***$_i$ maruthuvamanaikku
She her+dat health not good is told+ins she+acc hospital+dat
azhaittu cenraan.}MCL
taken went+3sgm
' He took her to hospital, because she said that she is not feeling well.'

3.1 {pookum vazhiyil}RP{ ***Siitaavin***$_i$ toozhi, Giita etiril vanthaaL}MCL
Going way+loc sita+gen friend geetha opposite came+3sgf
'On the way Sita's friend Gita came.'

3.0 {pinnar viiTiRku vantavuTan, ***avan***$_j$ ***avaLukku***$_i$ *roTTiyum paalum tayaar ceytu*
After home+dat came he she+dat roti+um milk made prepared
koTuttan.} MCL
give+3sgm
'After coming home he prepared bread and milk for her.'

In this example, in the 3.0 sentence there are two pronouns ***avan*** and ***avaLukku.*** Eventhough ***avaLukku*** refers ***Siitaavin***$_i$, Gita is the nominative proper noun which could be the most probable antecedent according salience factor. Still ambiguity remains whether the antecedent is Gita or ***Siitaavin***$_i$ **.** In sentence 3.2 *Raman* is a subject , which is dropped. Hence *avan*$_i$ in the sentence 3.0 which refers to *Raman* which is in the 3.3 sentence.

Example 4:

4.1 ***Raman***$_i$ aluvalagathil iruntu viTTiRku vanthu}NF {***kuzhantaikaLotu***$_j$
Raman office+loc from home+dat come children+soc
viLaiyaaTinaan.}MCL
played+sgm
'Raman came from office and played with children.'

4.0 {***avan***$_i$ ***avarkaLotu***$_j$ viLayaaTiviTTu} NF {***avarkaLukku***$_j$ inippu koTuttaan.}MCL
He their+soc played them sweet gave+3sgm
'After he played with them, he gave sweets to them.'

According to the agreement of person, number and gender, here in this example particularly the number i.e singular or plural is distinguished between avan and avarkal. And also in the above example if we replace kuzhantaikaL as kuzhantai, the gender, number variation is there. For **kuzhantaikaL** *– neuter, plural, 3 and its possible pronoun is avarkaL.*

For ***kuzhantai*** *– neuter, singular, 3 and its possible pronouns are atan, atu, atarku..*

3.2 Examples for 'atu'

Example 5:

5.1 {Naakaraajar muulastaanatil innum ***oru naakam***$_i$ uyiroTu irukkiRatu.} MCL
Naakarajar temple+loc still one snake alive be+3n.
'Still a snake is alive in Nakarajar temple'.

5.0 {***atu***$_i$ atikkaTi paktarkaLukku kaaTchi tarukiRatu}RP {enpatu kuRippTattakkatu.} MCL
It often deovotee+dat dharshan give+3sn is remarkable
'It is remarkable that it often gives its appearance to devotees'

In the above example the anaphor ***atu*** refers the antecedent which is animate and non-human living being, a noun phrase in the previous sentence.

Example 6:

{*piRaku poonkuzhali, "****cakravarttikku uTampu cukamillai enRu collkiRaarkale***$_i$,}COM
After poonkuzhali king+dat health not good be told+3pl
{***atu***$_i$ unmaitaanee?" enraal.}MCL
it true? said+3sgf
'After that Poonkuzhali said , "Everybody is telling king is not well", is it true?.'
Anaphor - **atu**$_i$, *Antecedent* - ***Immediate complement clause***

From the analysis of 'atu', it refers noun phrase such as place, object, animals, events or reasons, or sometimes behaves as deictic. ***atu*** refers noun phrases but sometimes it does not refers noun phrase instead it refers an event, reasons, a statement etc ... when it is followed by, atu+adverb, atu+postposition etc.

eg: **atu kaarNamaaka**
It because+ben
'Because of it'.

Similarly, some examples are, atu een?/ **It why?**, atu pola /**It like,** atu polave/**It like+e(clitic marker).**

From these examples we concluded that the antecedent is a noun phrase, or verb phrase or combination of both. The noun phrase kind of antecedents are either subject with case markers nominative or dative or possessive in Tamil. The noun phrase antecedents are taken for our work

to resolve. The analysis of ***'atu'*** reveals that its antecedents are mostly events or reasons which consists of noun phrase and verb phrase. Also it refers noun phrase. From the analysis, we found a set of features to build a system for anaphora resolution.

4 Our Methodology

In this task we have taken five sentences above from the anaphor occurred sentence and applied the salience measures and few other text information as our features to train our system. In machine learning technique feature identification is very vital part to give best systems. The features we used are discussed below.

4.1 Features for anaphora resolution

The corpus is manually analysed and features were identified. The features are classified into six categories such as word, POS, chunk, syntactic information, clause, Named Entity Recognition. The syntactic information consists of word category,gender,number,person. The detailed features given in train and test file are sentence recency, subject emphasis, object, proper noun, case of noun phrase, case of anaphor, current clause where anaphor occurs, immediate clause and non-immediate clause of anaphor of the sentence, whether a proper noun followed by a pronoun? and NE.

Since from our analysis, the pronominal *avan or aval* always refers some person, ultimately a proper noun. a*van or aval are* the pronoun which are always traverse backward from the current sentence and to other previous sentence to found the antecedent. During the traverse it may also refers the same person as pronoun like

(Antecedent) NNP ← NN (PRP) ← NN(PRP) ← PRP (anaphor)
Sita ← avaluteya ← avalukku ← aval
pacu/naakam/penna ← athanuteya ← atharkku ← atu
cow/snake/pen

Hence it is observed that subject or object which is noun or proper noun in the sentence or clause could be the most probable antecedents for the pronoun. From the analysis we can conclude that, in Indian languages, a nominative noun phrase, a possessive noun phrase with a nominative head and a dative noun phrase could be a subject of a sentence. So we have three types of subject nouns and in that, the most common the nominative noun. Hence nominative proper noun is given a very high score of 80 and the other two are given a score of 50 each. An NP with accusative case gets the next highest score of 40. NPs with other case markings (N. other) get a score of 30.The current sentence in which the pronoun occurs gets a score of 50 and this gets reduced by 10 for each preceding sentence.

S.no	Features	Definitions
1	Current Sentence	Sentence under consideration
2	Current Clause	Clause in which the anaphor occurs
3	Immediate clause	Clause next to the current clause
4	Non-Immediate clause	Neither an immediate nor a current clause
5	NNP/NN-Nom	Any proper noun/noun nominative

6.	NNP/NN-Poss	Any proper noun/noun possessive
7	NNP/NN-Dat	Any proper noun/noun Dative
8	NNP/NN-acc	Any proper noun/noun accusative
9	NNP followed by any pronoun with or without case marker?	NNP followed by immediately by pronoun with any of the case marker?

Table 1 – Features

The example 7 illustrates the case where the pronominal occurring after the proper noun which is the subject of the sentence cannot have it as the antecedent for the pronominal. Only if such a pronominal is reflexive then it is possible to have the subject as the antecedent. This rule has exceptional case where the pronominal *avanai* or *avalai* acts as a reflexive. Example 8 illustrates the exceptional case.

Example 7:

KrishNan ***avanai*** azhaitaan.
Krishnan he+acc called+3sgm
'Krishnan called him'

In this example ***avanai*** it never refers Krishnan, an immediate pronoun.

Example 8:

KrishNan avanai maaRRikoNTtaan
Krishnan he+acc change+do+3sgm
'Krishnan changed himself.'

In the case of reflexive sentence as in the above example, KrishNan is a nominative proper noun which is followed by anaphor ***avanai*** with accusative case marker**,** refers KrishNan. Here krishnan talks about himself. So here avanai refers KrishNan.

5 Corpus

The corpora we have considered for this work is from the web. We have taken the web data from tourism domain which consists 10000 sentences. Annotation guidelines are formulated and tagged with the following guidelines. To annotate the corpus with anaphor and antecedents with index, we used an annotation tool, PALinkA. We have considered both anaphor and antecedent as markables. For annotations, first anaphor and antecedent should be marked as markables and if it is anaphoric, link is established between these two markables. Finally all the possible anaphor and antecedents are tagged with index.

6 Machine Learning Algorithm

6.1 Conditional Random Fields

CRF++ is toolkits designed for generic purpose and are applied to a variety of NLP tasks. The machine learning method of CRFs was chosen to do our experiments, because of its flexibility to build linguistic rules. CRFs can contain number of feature functions. The advantage

of CRFs is that it can model not only sequential data, but also non-linear data. Since the task of extraction of antecedents is syntactic and semantic task, CRFs is appropriate for this purpose.

Our Algorithm:
Step1: Input the corpus.
Step2: Preprocess the corpus, tagged with POS, Chunk, word category, gender, number, person, suffix, case markers, Clause Information and Named Entity Recognition.
Step3: Tag the Anaphor and the Antecedent using PALinkA tool.
Step4: Identify the features and given to CRF.
Step5: Train and test both the Train corpus and Test corpus with the features identified in CRF.

7 Result and Discussion

7.1 Ten-fold Experiment:

We have taken the tourism corpus from web of 10000 sentences. The total numbers of words are 93102 and we split this word corpus equally into 10 equal files. Randomly we have taken 8 files for Training (74482 words) and 2 (18620 words) files for testing with all combinations of files. The first corpus consists of 945 pronouns. The training data consists of 829 pronouns and the testing data consists of 116 pronouns. The results are given below.

S.no	Total no anaphora in training	Total no anaphora in testing	System tagged	System tagged correctly	Recall	Precision
1	794	149	147	123	88.55%	83.67%
2	769	174	173	140	80.45%	80.92%
3	782	171	168	142	83.04%	84.52%
4	798	145	143	122	84.14%	85.31%
5	766	177	176	140	79.09%	79.84%
6	652	291	289	242	83.16%	83.73%
7	613	330	318	261	79.09%	82.07%
8	738	205	196	162	79.02%	82.65%
9	827	116	112	91	78.44%	81.25%
10	815	128	123	99	77.34%	80.48%
				Average	80.63%	82.44%

Table 2 – Tenfold Experiment results

8 Conclusion and future work

This work considered all kind of pronominals for resolution. This work requires further analysis and fine tuning of the features. Still there are lot of challenges in anaphora in discourse structure. And resolution of pronominals is very much needed for all natural language processing applications.

References

Carter, D.(1987) *Interpreting anaphors in natural language texts*. Chisester: Ellis Horwood ltd.

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

10. 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.

Jha, G.N., Sobha, L, Mishra, D., Singh, S.K., Pralayankar, P. (2008) *Anaphors in Sanskrit* In: Proc. Second Workshop on Anaphora Resolution Johansson, C.(Ed.).

Dhar, A. and Garain, U. (2008) *A method for pronominal anaphora resolution in Bengali* In: proc. 6th Int. Conf. on Natural Language Processing (ICON) at Pune, India, December.

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

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.

Lafferty, J., McCallum, A. and Pereira, F. (2001) *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*. In: 18th International Conference on Machine Learning, pp .282--289. Morgan Kaufmann, San Francisco, USA.

Murthi, N.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.

Orasan, C. (2003) *PALinkA: a highly customizable tool for discourse annotation.* In: proc. 4th SIGdial Workshop on Discourse and Dialog, Sapporo, Japan, 5 – 6 July, pp. 39 – 43.

Sobha, L., Patnaik, B.N. (1999) *VASISTH- An Anaphora Resolution System Unpublished Doctoral dissertation*. Mahatma Gandhi University,Kottayam,Kerala.

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

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

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