A Semi-automated Annotation of Co-reference Chains in Tamil

Vijay Sundar Ram R and Sobha, Lalitha Devi

AU-KBC Research Centre
MIT Campus of Anna University, Chennai-600044

sobha@au-kbc.org

Abstract. Co-referential chains are formed by grouping various anaphoric expressions referring to the same entity. We present our co-reference annotation schema for Tamil text. Co-reference annotation guidelines is less attempted in Indian languages. The annotation guidelines is designed for annotating various anaphoric expression such as pronominals, reflexives, reciprocal, distributives, one anaphor, noun-noun anaphora and definite descriptions. As PRO-drop is common across Dravidian languages, we identify zero pronouns and handle them using a semi-automated method before annotating the corpus. The semi-automated method includes rule-based algorithm for zero pronoun identification and manual verification. In this paper we have discussed about the schema for annotation of co-refering entities and the various statistics related to the annotated corpus. A detailed study on inter-annotator agreement of the annotated corpus is presented.

Keywords: co-reference annotation, Tamil, semi-automatic, PRO-drop

1 Introduction

Machine learning technique requires large quantity of annotated data for training. Co-reference annotated corpora are available in languages such as Catalan, Dutch, English, German, Italian, Arabic, Spanish, and Polish etc. There is no Co-reference annotated corpus available in Indian languages. Though NLP Tool Contest in ICON2011 had a shared task titled, ‘Anaphora Resolution in Indian Languages’ had anaphora annotated corpus in three Indian languages, did not have co-reference annotation. Also among the anaphors only pronominals with their antecedents annotation was dealt. And these annotated corpuses were available for three different languages namely, Tamil, Hindi and Bengali. Due to the non-availability of co-reference annotated data for Tamil, preparation of annotated corpus became inevitable.

Co-reference annotation in Indian languages is less attempted, which is the major reason for lack of this annotated corpus is non-availability of annotation guidelines for Indian languages and very less research in this field. The annotation of corpus requires proper guidelines, which define the tasks unambiguously. An unambiguous guideline helps to create an annotated corpus with a high degree of inter-annotator agreement. There are annotation guidelines for co-reference annotation for various languages such as English, Spanish, Catalan, Arabic, Chinese, Italian, Polish, etc.

One of the earliest annotation guidelines for co-reference annotation is Message Understanding Conference (MUC) standards presented by Hirschman [1] and DRAMA schema by Passonneaus [4]. DRAMA schema includes instruction for dealing with markables annotation in dialogues and is for monolingual. MUC Coreference standard by Hirschman [1] was designed with importance to preserve high-annotator agreement than to capturing every possible phenomenon that could fall under co-reference. This annotation schema covers only Identity (IDENT) relation of the noun phrases. It does not include co-reference relations among clauses and set/subset, part/whole relation. The IDENT relation was defined to be symmetrical and not directional. Verb co-reference is not attempted. This schema does not provide instructions to annotate zero pronouns and empty strings as markables. Appositive, metonymy and predicate nominals are considered to have IDENT relation. MUC schema does not provide instructions to annotate co-referential entities in dialogues. It was designed for English and therefore did not include instructions for anaphoric expressions common in other European languages such as clitics, elliptical pronouns etc.

MATE Annotation schema was proposed by Poesio et al. [5]. The goal was to develop a schema to annotate co-reference at different levels that could be useful to different types of applications. It has core schema and three extensions. The core schema can be used to annotate the type of annotations that can be done with MUC schema. The three extensions to the core schema can be used to annotate i) reference in visual situations, ii) more complex set of relation between entities iii) anaphoric relation involving an extended range of anaphoric expressions. This schema only recommends making potential antecedents of anaphoric and referential expressions that can be realised in full NPs. MATE annotation schema can be used for annotating dialogues and it also provide instructions to annotate clitic pronouns, elliptical pronouns and split antecedents. In MUC schema partial NPs can be annotated. One of the well-known annotation schemas is Automatic Content Extraction (ACE) schema which was developed for ACE 2005 and ACE 2007 shared task, where the entities and its relations have to be detected. Here the types of entities are defined and the relations are marked between these entities.

Most of the works in noun phrase co-reference were restricted to named entities due to lack of availability of general anaphoric co-reference data. As mentioned above ACE data set was restricted to a set of named entities. To overcome this, OntoNotes project aimed to create a large-scale accurate corpus for general anaphoric co-reference that covers events, entities and not limited to noun phrases. In OntoNotes, two different types of co-reference relations namely Identical (IDENT) and Appositive (APPOS) were used. Here nominal predicates are not considered as co-referential [6].

Recasens et al. [7] has presented AnCora-Co, Spanish and Catalan co-reference annotated corpus. In this annotation, they have dealt with the annotation issues in Elliptical pronouns, Clitic pronouns, Quoted speech, Possessives, Embedded NPs, Split antecedents, Referential vs attributive NPs, Generic versus Specific NPs, Metonymy and Discourse Deixis. This co-reference annotation includes entities and events. The types of co-reference relations used in AnCora-Co are Identity, Discourse Deixis, and Predicative.

Further the paper is presented as follows. In section2, we have discussed our annotation schema, types of markables, types of co-reference relations between the markables. Genres of the corpus, steps followed in annotation task and identification of Zero pronouns are presented in section 3. Corpus statistics is presented in section 4. Inter-annotator agreement in the annotated corpus is presented in section 5. The paper concludes with a summary of the work done in conclusion section.

2 Annotation Schema

In this section, our annotation guidelines is explained in detail. The co-referential entities relations focussed here are the following: Identity relation, and Definite Description relation. The span of annotation is full NP and not partial NP, this is similar to MATE annotation schema. The annotation schema includes split antecedents One-Anaphors and has not included clauses as markables, whole/part relation, hyponym, meronym and metonym relations. In the following sub-sections, we describe the types of markables (referential entities) and the relations between co-referential entities.

2.1 Types of Markables

In co-reference annotation, the members of the co-reference chains are called as Markables. The types of Markables included in our annotation task are as follows.

1. Pronominal: This includes1st, 2nd and 3rd Person singular/plural Pronouns such as naan (I), engaal (our), nee(you), ungaL (your), avan (he), avaL (she), athu (it), avai (they).
2. Reflexive Pronouns: This includes reflexives, such ‘avane’ (himself), ‘avaLe’ (herself), whose antecedent will be the subject in the same clause.
3. Reciprocal Pronouns: ‘oruvarukkoruvar’ (eachother)
4. Distributive Pronouns: ‘avaravarukku’ (their)
5. Proper Nouns: This includes entities such as Person, Place, Organisation, Artifacts etc.
6. Demonstrative Nouns: This includes noun phrases with demonstrators such as ‘anthakuuttam’ (that meeting), ‘ikkovil’ (this temple).
7. Definite Description Nouns: These are the noun phrases, which occur preceding or following the entities. These are the denoting phrase of an entity are such as Chief Minister, Captain, etc.
8. Cardinal Noun such as ‘onru’, (one), ‘iraNtu’ (two), is considered as noun phrase instead of quantifiers.

2.2 Types of Co-referential Relation

Anaphoric Relation.

The relation between the anaphors and their antecedents are annotated with different types of anaphoric relations based on the type of the anaphora. The types of the anaphors are as follows: pronominal, noun-noun anaphoric relation, reflexives, reciprocals, distributives, and One anaphors.

Definite description.

Relation between the denoting phrase of an entity and the entity is annotated with Definite Description relation. Consider the following examples Ex.1.

Ex.1.a

Pirathamar	moodi
Prime Minister(N)	Modi(N)

Ex.1.b

raman	daktar
Raman(N)	Doctor(N)

In Ex.1a, ‘pirathamar’ (Prime Minister) is the definite description of Noun phrase “Modi”.

Ex.1.b has a copula drop. Here the definite description ‘daktar’ (Doctor) follows the noun phrase ‘Raman’, a person entity.

3 Corpus Annotation

We aimed at developing an annotated corpus for building a robust co-reference resolution engine which suits for various Tamil web-contents. Hence we collected News articles from various online Tamil News wires. The News articles are from Sports, Disaster and General News domains.

We have followed a semi-automated methodology to annotate co-referring entities in Tamil text. We started with preprocessing the text with syntactic modules, namely, Morphological analyser, POS tagger, Chunker, Clause Boundary identifier. Then we further processed the text with Named Entity Recognizer. The syntactic information enriched text is processed with a rule-based engine to identify the zero pronouns in the text. The pronouns introduced in the PRO-drop positions are manually verified.

The anaphoric expressions are annotated along with its antecedents usin graphical tool, PAlinkA, a highly customisable tool for Discourse Annotation (Orasan, 2003). We have used two tags namely, MARKABLE and COREF.

For anaphoric pronoun its possible antecedent, noun phrasre, can occurs in the same sentence preceding to the pronoun or in the preceding sentences. The anaphoric pronoun and its antecedent match in person, number and gender. 3rd person neuter pronoun ‘atu’ (it) can have a clause or sentence as an antecedent, where an event is described. In the present work we have not handled clause or sentence as antecedent.

The antecedent of reflexive is the subject in the same clause. For reciprocals and distributives, possible antecedent will be a plural noun phrase which match in person, number and gender and occurs in the same clause. The antecedent of the One-anaphor will be a non-nomative noun phrase with quantifier preceeding the head noun and occur in preceding clause or sentence.

In a text, a noun phrase may be repeated as a full noun phrase, partial noun phrase, acronym, or semantically close concepts such as synonyms or superordinates. These noun phares are annotated as noun-noun anaphor and antecedents. These noun phrases mostly include named entity such as Individuals, place names, organisations, temporal expression, abbreviation such as ‘juun’ (Jun), ‘nav’(Nov) etc., acronyms such as ‘i.na’ (U.N), etc., demonstrative noun phrases such as ‘intha puththakam’ (this book), ‘antha kuuttam’ (that meeting) etc., and definite descriptions such as denoting phrases.

Definite Descriptions is a unique denoting phrase of an entity. These phares occur immediately preceding the entity or following the entity.

Identification of zero pronouns in Tamil text is explained in the following sub-section.

3.1 Zero Pronoun Identification

In certain languages, the pronouns are dropped when they are grammatically and pragmatically inferable. This phenomenon of pronoun drop is also mentioned as ‘zero pronoun’, ‘null or zero anaphors’, ‘Null subject’. This phenomenon is also common across Dravidian languages. Consider the discourse in Ex.2, which has two sentences. These two sentences have focus on the same entity.

Ex.2.a

thalaivar	*kuRiththa*	*neraththil*	*vizaviRku*	*vanthaar.*
The leader(N)	exact (RP)	time(N)+loc	function(N)+dat	come(V)+past+3sh

(The leader came to the function at exact time.)

Ex.2.b

PRO	*thaane*	*kaarai*	*ootti*	*vanthaar.*
He(Pn)	himself(reflexive)	car(N)+acc	drive(V)	come(V)+past+3sh

([He] himself drove the car and came.)

Ex.2.c

PRO	thaane	kaar	ootta	veNtum	enRu	aacai.
He(Pn)	himself(reflex)	car(N)+nom	drive(V)+inf	want(V)	that(comp)	like(V)

([He] liked to drive the car himself.)

In Ex.2, the second sentence Ex.2.b has a reflexive and the reflexive always refer to the Subject noun in same clause. Here in this sentence the Subject is dropped. The dropped subject noun is a nominative noun. In Ex.2.c, the subject NP, which is the antecedent of the reflexive ‘thaane’ is dropped. The finite verb of this sentence, ‘pitikkum’ (likes), which conveys the semantic notion of ‘liking’, so the subject NP of this sentence should have dative case marker.

Zero pronouns are common in complimentizer clause sentences. Following examples Ex.3, Ex.4 and Ex.5 explain, zero pronouns in complimentizer clause.

Ex.3.a
naan sithaavin thanthaiyai kaNteen.
I(PN) Sita(N)+gen father(N)+acc see(V)+past+1s
(I met Sita's father.)

Ex.3.b
sithaa naalai varuvaaL enru PRO kuRinaar.
Sita(N) tomorrow(N) come(V)+past+3sf that(comp) he say(V)+past+3sh
(He said that Sita will come tomorrow.)

Ex.4.a
naan sithaavin thanthaiyai kaNteen.
I(PN) Sita(N)+gen father(N)+acc see(V)+past+1s
(I met Sita's father.)

Ex.4.b
sithaa naalai varuvaaL enRaar.
Sita(N) tomorrow(N) come(V)+past+3sf that(comp) +(he said)
(He said that Sita will come tomorrow.)

Ex.5.a
naan sithaavin thanthaiyai kaNteen.
I(PN) Sita(N)+gen father(N)+acc see(V)+past+1s
(I met Sita's father.)

Ex.5.b
avar, sithaa naalai varuvaaL enru kuRinaar.
He, Sita(N) tomorrow(N) come(V)+past+3sf that(comp) say(V)+past+3sh
(He said that Sita will come tomorrow.)

Examples Ex.3, Ex.4 and Ex.5 have a discourse of two sentences and all these examples have same sentences written in different styles. In Ex.3.b, the main clause is only the verb phrase 'kuRinaar' and the subject is dropped, giving rise to a zero pronoun. In Ex.4.b, the complimentizer is frozen with the verb phrase 'enraar'. Here 'enraar' has occurred in the place of 'enru avar kuRinaar' (that he said). Here the pronoun is not explicit and it occurs as a zero pronoun. Ex.5.b has a clausal structure where the complimentizer clause is embedded within the main clause.

We have attempted to identify zero pronouns in the sentences with reflexives and also in sentences with complimentizer clause. We have used a rule based engine to identify zero pronouns in two sentence structures. The algorithms are described below.

Algorithm to identify zero pronouns in sentences with reflexives

Step 1: Check for reflexive pronoun in the sentence. If YES, go to step 2.

Step 2: Check if the finite verb is cognitive verb or not. If exist, and if the subject NP is in dative case, then go to step 3 else step 4.

Step 3: Check for NPs/ possessive NPs having head noun with dative case in the sentence, if NO, go to step 5 else exit.

Step 4: Check for NPs/ possessive NPs having head noun with nominative case in the sentence, if NO, go to step 7 else exit.

Step 5: Extract PNG information of the finite verb.

Step 6: Introduce pronoun based on the PNG information with dative case in the beginning of the clause.

Step 7: Extract PNG information of the finite verb.

Step 8: Introduce pronoun based on the PNG information with nominative case in the beginning of the clause.

Algorithm to identify zero pronouns in complimentizer clause sentence

Step 1: Check for sentence with complimentizer clause.

Step 2: If the sentence do not have complimentizer clause embedded in the main-clause, go to step 3.

Step 3: If the main clause has enraar/enraaL/enrathu, replace it with ‘enru avar kuRinaar’/ ‘enru avaL kuRinaaL’ / ‘enru athu kuRiyathu’.

Else

Check for nominative noun phrase in the main clause. If does not exists then go to step 4.

Step 4: Extract PNG information of the finite verb in the main clause.

Step 5: Introduce a nominative pronoun on based on the PNG information in the subject slot of same clause.

4 Corpus Statistics

In this section, we present various statistics of the annotated corpus. The basic statistics of the annotated corpus is presented in table 1.

Table 1. Basic Corpus Statistics

Details about Corpus	**Count**
Number of Web Articles annotated	1,000
Number of Sentences	22,382
Number of Tokens	272,415
Number of Words	227,615

Details on the distribution of the anaphoric expressions are presented in the following table 2.

Table 2. Distribution of anaphoric expressions in the Corpus

S.No	**Type**	**Number of Occurrence**
1	Noun-Noun Anaphora	11,935
2	Anaphoric Pronominal	4,160
3	Definite-Description	1,890

4	Reflexives	29
5	Reciprocal	31
6	Plural pronouns with split-antecedent	190
7	Distributives	8
8	Zero Pronouns	453
	Total	18,696

5 Inter-Annotator agreement

Inter-annotator agreement is the degree of agreement among annotators. It is the percentage of judgements on which the two analysts agree when coding the same data independently. There are different statistics for different types of measurement. Some are joint-probability of agreement, Cohen's kappa and the related Fleiss' kappa, inter-rater correlation, concordance correlation coefficient, Cochran's Q test, intra-class correlation and Krippendorff's Alpha. We use Cohen's kappa as the agreement statistics. The kappa coefficient is generally regarded as the statistics of choice for measuring agreement on ratings made on a nominal scale. It is relatively easy to calculate, can be applied across a wide range of study designs, and has an extensive history of use.

The kappa statistics K is a better measure of inter-annotator agreement whichtakes into account the effect of chance agreement [2].

$$K = (p_0 - p_c)/(1- p_c)$$

where p0 is agreement rate between two human annotators and p_c is chance agreement between two annotators.

The results of kappa-like agreement measurements are interpreted in six categories as follows (Yalçınkaya et al. 2010).

1, Measurement> 0.8: Perfect agreement
2, 0.8 >Measurement> 0.6: Substantial agreement
3, 0.6 >Measurement> 0.4: Moderate agreement
4, 0.4 >Measurement> 0.2: Fair agreement
5, 0.2 >Measurement> 0.0: Slight agreement
6, 0.0 >Measurement: Poor agreement

The annotation of anaphoric expressions was done by three annotators. W calculated the kappa score for each type of anaphoric expressions and it is presented in the following table 3.

Type	**Kappa Score (K)**
Pronominal	0.79
Reflexives	0.86
Reciprocals	0.89
Distributives	0.91
Noun-Noun Anaphora	0.73

Definite Description	0.65

Table 3. Kappa Scores for DifferentAnaphoric Expressions

The overall Kappa score (K) is 0.78. On analysing the kappa scores of Reflexives, Reciprocals and Distributives respectively it is found that they have perfect agreement.

The difference between the annotaters were analysed and found the variation in annotation. It occurred in the marking of antecedents for pronominal. This is common in sentences with clausal inversion, and genitive drop. Consider the following discourse Ex.6.

Ex.6.a
raamu naaLai varuvaan enRu coomu
Ramu(N) tomorrow(Adv) come(V)+future+3sm that(comp) Somu(N)
connaan.
say(V)+past+3sm
(Somu said that Ramu will come tomorrow.)

Ex.6.b
avanukku raamuvin thampi kuuRinaan.
He(Pn)+Dat Ramu(N)+gen brother(N) say(V)+past+3sm
(Ramu's brother said to him.)

For the discourse in Ex.6, two annotators have wrongly annotated 'raamu' (in Ex.6.a) as the antecedent of the pronoun 'avanukku' (to him) in Ex.6.b. And one annotator-correctly tagged 'coomu' as the antecedent. The confusion has occurred because of the clause inversion in Ex.6.a.

There are confusions in antecedent of reflexives when the subject is dropped (zero-pronoun) which cannot be correctly identified in the sentence.

Inter-annotator agreement is relatively low in the annotation of entities and Definite Description. The reason for this confusion is high occurance of genitive drop in Definite Descriptions. Consider the following example 7.

Ex.7
then maNtala varththaka maiya thunai thalaivar
Southern(adj) Regional(N) Trade(N) Centre(N) Deputy(N) Head(N)
thiru. cukumaar.
Mr.(N) Sukumar (N)
(Southern Regional Trade Centre Dupty Head Mr. Sukumar)

In Ex.7, 'then maNtala varththaka maiya thunai thalaivar' is the Definite description of personentity 'cukumaar'. The DD has three NPs namely 'then maNtala' 'varththakamaiya' and 'thunai thalaivar', but the first two NPs has genitive drop, which leads to confusion.

In Noun-Noun anaphora, there are confusions are due to spell variation, aglization, synonymous usage for the sameentities. 'ciRaiathikaari' (prison officer) and ‘ciRaikaavalar’ (prison police) refer to the sameentity but the synonymswords are used. Similarly in the case of place names, the News articles will have place names and description referring to the place name. These bring in consistencies in tagging-these nouns. Consider (Ex.8).

Ex.8.a
cennai
Chennai(N)

Ex.8.b
tamizaka thinthalai nagaram
Tamil_Nadu’s Capital

Ex.8.a and Ex.8.b both refers to the same entities. The first one is the actual place name and the second one is the description referring to the place name.

6 Conclusion

We have described a semi-automatic method for annotation of co-reference chains for Tamil text. We have discussed on the statistics and inter-annotator agreement of the annotated corpus The salient points presented in the paper are as follows:

- Types of markables in our annotation schema included are Pronominals, Reflexives, Reciprocals, Distributives, Proper nouns, Demonstrive nouns, Definite Descriptions and Cardinal nouns.

- Types of co-referential relations used in our annotation schema are anaphoric and Definite Description.

- Zero pronouns are identified automatically using rule-based engine by processing the information enriched text and manually validated.

- Corpus, which is used for annotation are collected from various online Tamil Newspapers. We have annotated 1000 News articles.

- Inter-annotator agreement was measured for each of the anaphoric expression. The major observations are the following. a) In sentences with clausal inversion and genitive drop there were disagreements in the annotation of antecedents of the pronominals. b) There were disagreement in antecedent of reflexives when the subject drop (zero pronoun) is not correctly identified in the sentences. c) The disagreement is high in the annotation of entities and their definite descriptions. The major reason for this disagreement is very high genitive drop. The disagreement of annotation in Noun-Noun anaphors is due to spell variation, aglization, synonymous usage for the same entities.

References

1. Hirschman, L.(1998). MUC-7 coreference task definition, version 3.0. In N. Chinchor, editor, In Proc. of the 7th Message Understanding Conference.

2. Ng, H.T, Lim, C.Y.and Foo, S.K.(1999). ‘A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation’, In Proceedings of the {ACL} {SIGLEX} Workshop on Standardizing Lexical Resources {(SIGLEX99)}, Maryland, pp. 9-13.
3. Orasan, C.(2003).PALinkA: A highly customisable tool for discourse annotation. SIGDIAL Workshop, pp. 39-43.
4. Passonneau, R.J.(1997). Instructions for applying discourse reference annotation for multiple applications (DRAMA)’, Unpublished manuscript
5. Poesio, M. and Kabadjov, M.(2004). A general-purpose, off-the-shelf anaphora resolution module: implementation and preliminary evaluation’, In: Language Resources and Evaluation Conference.
6. Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R.&Xue, N.(2011) ‘Conll-2011 shared task: Modeling unrestricted coreference in ontonotes’, In Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011), Portland, Oregon.
7. Recasens, M.and Marti, M.A. (2010). AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan, Language Resources and Evaluation, vol. 44, no. 4, pp. 315-345.
8. YalçınkayaIhsan, S. (2010). An Inter-Annotator Agreement Measurement Methodology for the Turkish Discourse Bank (TDB), A thesis submitted to the Graduate school of informatics of the Middle East Technical University.