Research Paper

ANAPHORA RESOLUTION IN TAMIL THROUGH DECISION TREE LEARNING - CHALLENGES AND TACKLES

Arul Deepa K[*], Deisy C

Address for Correspondence

Department of Information Science and Technology, College of Engineering, Guindy, Anna University, Chennai – 600 025, India
Department of Computer Science and Engineering, Thiagarajar College of Engineering, Madurai - 625 015, India

ABSTRACT:

This manuscript presents a decision tree based machine learning algorithm for Tamil Anaphora Resolution. For the received input Tamil text document the system first identifies mentions which are pronouns (of the type personnel, reflexive, possessive and distributive) and noun phrases. Then characterization of each mention is done by a pipeline of dispensation namely segmentation, rule-based morphological analysis and noun phrase chunking. This system addresses the additional challenges of Tamil like morphological richness, semantic ambiguity and more deictic. Then the system built a feature vector of lexical, syntactic and semantic features for machine learning. The preferred decision tree based system has obtained f-measure of 73% and up to 77% accuracy, which is an encouraging result for a challenging language like Tamil. The experiments were done on NER corpus (Tourism) of TDIL dataset, which encourages future comparisons against the obtained results.

KEYWORD: anaphora resolution, decision tree, machine learning, pronoun resolution, Tamil computing.

1. INTRODUCTION

Entities in natural language text are not recognized with handy unique ids, but quite with frequently varying metaphors. For an example,

பெரியார் என்று பரவலாக அறியப்படும் ஈ. வெ. இராமசாமி சமூக சீர்திருத்தத்திற்காக போராடியவர். இவர் திராவிடர் கழகத்தினைத் தோற்றுவித்தவர். இவருடைய சுயமரியாதை,இயக்கமும், பகுத்தறிவுவாதமும் மிகவும் புகழ்பெற்றது.

Transliteration: "Periyaar enRu paravalaaka aRiyappadum I. ve. Iraamasaami samuuka siirthiruthathRkaaka pooraadiyavar. ivar thiraavida kazakaththinaith thooRRuviththavar. Ivarudaiya suyamariyaathai iyakkamum, paguththaRivu vaathamum migavum pugazpeRRathu.

Translation: E. V. Ramasamy, affectionately called Periyar was a social activist. He was the founder of Dravidar Kazhagam political party. His self-respect movement and rational propagation are very famous.

Here 'Periyaar' (Periyar), I. ve. Iraamasaami' (E.V. Ramasaamy), 'ivar'(He) and 'ivarudaiya' (His) are the descriptions which downgrade the alike element. Here is the undertaking to develop a life form for programmed gathering of every single composed descriptor that refer to the same substance inside of a coreference chain either intra or entomb records. The issue of mapping phonetics expressions into these hidden elements is known as reference determination. Anaphora is the reference to an element that has been already brought into the discussion. It depicts an awry connection between two semantic expressions, an anaphor and an antecedent where, anaphor (இவர்/he) is the alluding expression and antecedent (பெரியார் /Periyaar) is the one being alluded [7]. Both corefer one another. Reference determination assumes a crucial part in numerous NLP applications including information extraction, question answering, automatic summarization, opinion mining, topic identification, machine translation and natural language generation. Thus resolving the references in natural languages is another attention needed area to be concentrated for the NLP research community. Apart from English this issue had been addressed in many languages like Arabic [9,10,21,22], Chinese [5], Hindi [8,16,18], Japanese[11,12,17], Tamil [1,3,4,15] and Turkish [13].

Based on the altitude of automation reference resolution approaches are classified as knowledge based (60s-80s), heuristic based (90s) and Machine learning based approaches. In recent times, reference resolution research has given the focal point in the direction of statistical and machine learning approaches [6,7,8,10,11,14,15,19,20] due to their domain independence and extendibility to various languages.

2. Literature Review

McCarthy and Lehnert [14] built up the primary machine learning approach, the Mention-Pair model based decision tree algorithm trained with MUC-5 English Joint venture corpus on every conceivable pairs demonstrated an outcome superior to the heuristic methodologies. All feasible markables in a training file are decided by a pipe-line of language processing modules, and training examples within the form of characteristic vectors are generated to suitable pair of markables. These training examples are then given to a learning algorithm to construct a classifier. For scan report all markables are decided and talents pairs of coreferring markables are provided to the classifier, which decides whether or not the 2 markables virtually corefer. This model was further extended with a set of twelve elements with the aid of quickly et al. [19] which can be a mention pair model that classifies whether or not two mentions are coreferent or not. Hindrance with their approach is its inadequate knowledge to make an instructed coreference choice and each candidate antecedents is regarded independently of the others. These models are referred to as single candidate model where a candidate is the antecedent of an anaphor. This only considers the candidates of an anaphor in isolation, which is incapable of simply shooting the option relation between candidates for its training.

Subsequent is twin candidate model [7, 11] recast anaphora resolution as alternative classification challenge that determines the choice between two competing candidates for the antecedent of a given anaphor. That is positive and negative instances are created which makes the training even better. Denis [7] came up with a ranking model where the candidates are regimented based on their fondness and the superlative the antecedent.

Yang et al. [20] explained entity point out model with multiplied expressiveness, that identifies whether a mention and a preceding probably partially formed cluster are coreferent or no longer. This makes it possible for computation of cluster stage elements. Concern of this method is each and every candidate cluster is regarded independently of the others. Dennis [7] proposed a mention rating model that imposes a ranking on a collection of candidate antecedents. Parallelism is its strength, i.e. this considers the entire candidate antecedents simultaneously. Primary crisis of such units is that they accumulate inadequate expertise to make an instructed coreference selection. Clustering based (unsupervised) systems are chosen for the reason that they take a extra world view of constructing coreference chains and are based both on constraint propagation or on probabilistic strategy [6].

Narayana Murthi et al. [15] proposed a salience measure process and computing device learning based algorithm for pronominal resolution in Tamil. This paper introduces a suite of nice tuned salience explanations and weights that extra suitable for the fairly free phrase order languages. The system is experimented with a corpus from CIIL (central Institute of Indian Languages, Mysore, India) and got 86.32% of precision and 80.9% of do not forget charges for more than a few circumstances of nominative, accusative, genitive, dative, locative and instrumental. Akilandeswari et al. [1] proposed a computer studying process for probably the most long-established pronominal of Tamil discourse-avan, aval, athu and their suffixes. The 5 classes of features are phrase, POS, chunk, syntactic understanding, clause and NER information. These points had been utilized to develop the method using conditional random discipline (CRF) model. The procedure was once confirmed with tourism area information of ten thousand sentences and showed encouraging outcome. A technique to get to the bottom of anaphors of persons, places, plurals and movements in Tamil textual content was proposed via Balaji et al. [4]. Many of the earlier systems had been headquartered on syntactic competencies of the language; right here is an method that incorporates semantic expertise (with the aid of using common Networking Language) additionally. Their triggering tuple approach works with discount of ambiguity whilst resolving. This method is evaluated with a corpus from tourism and news domain and completed 84% accuracy and is when compared with the earlier rule headquartered method.

3. Overview of Tamil Reference System

Like different Indo-Dravidian languages Tamil is also an agglutinative language. Tamil words consist of a lexical root with a number of affixes connected. As a rule by using examining the morphemes of a noun phrase (NP) person, number, gender and lots of like agreement elements will also be recognized. Tamil does no longer have a definite article. So the characteristic, specific pronoun agreement between the anaphor and antecedent cannot ever be recognized. Tamil has character, number and gender contract. Ideas of Tamil Language have many clues to symbolize the mentions. Each Tamil noun ends with any of those distinct characters (ஆ-A, இ-i, ஈ-i, உ-u, ஊ-uu, ஐ-ai, ண்-N, ம்-m, ய்-y, ர்-r, ல்-l, ழ்-z, ள்-L, ன்-n) [2].

3.1. Tamil Pronouns

Tamil pronouns are classified as private, integrative, reflexive, possessive, demonstrative and distributive. There aren't any relative pronouns, instead participles are used. Additionally there's no direct pronoun list for possessive category; alternatively the genitive circumstances of alternative pronouns are categorized as possessive ones. Summary of Tamil pronoun method with examples is defined in figure 1.

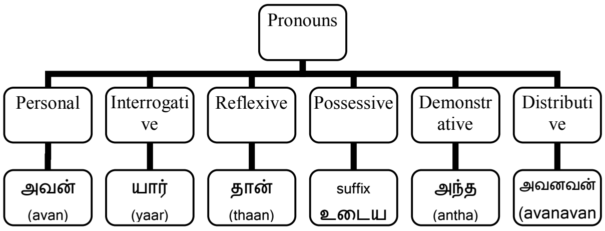

Fig. 1. An overview of Tamil pronoun System

4. General Algorithm

Layout of the strides remains nearly the same for any language considered and whatever may be the methodology selected, the outline or order of the steps remains almost the same [7]. The method for these strides are being completed relies on upon the technique and nature of the dialect. The rearranged adaptation of the general algorithm is displayed in figure 2.

1. Mention Detection & Characterization	Identifies list of markables and characterize
2. Anaphoricity Determination	Identifies based on linguistic Characteristics
3. Antecedent Generation	Identifies a list of possible antecedents
4. Pair/Rule Formation	Makes all possible {anaphor, antecedent} pairs
5. Filtering	Reduces the size of antecedent list
6. Scoring / Ranking	Ranks or assigns salience values
7. Searching / Clustering	Decides the best matched antecedent

Fig. 2. General Algorithm

5. System Design

In this section we introduce the engineering of the proposed System of Tamil Anaphora Resolution (STAR) and clarify different challenges confronted at every module. This setup is done in two folds. To begin with is from the dialect information standards were recognized to fabricate a successful component vector and the second is the working of classifier to get the framework learn. The following subsections explain actions carried out to build an effective feature vector and the classifier module of the experiment.

5.1. Processing Pipeline

According to Soon et al. [19], a pipeline of processing like segmentation, POS tagging, Named Entity recognition are needed to identify and characterize the markables. There are two ways to achieve this. One is by following appropriate tool and the other is doing the work with tagged dataset. As it is a great deal to develop perfect tools for Tamil language (A 100% perfect POS tagger is still a challenge even to English), we considered a tagged corpus with the assumption that the annotation is 100% right.

5.2. Markable Detection

Collections of pronouns that need to be resolved are called anaphors and collection of noun phrases and nested noun phrases are called as antecedents. Union of anaphors and antecedent are known as markables, which are identified by tag analysis. Some way of pre-processing is needed for better handling of them. There are two cleaning steps involved. One is the chunking of noun phrases.

Example:தமிழ்நாடு+சுற்றுலா+வளர்ச்சிக்+கழகம்

Transliteration:

tamilwaadu+sutRRulaa+vaLarssik+kazakam

Translation:

Department of Tamilnadu Tourism-Development

Second step is stemming, done in two levels. First is stemming of pronoun extensions.

Example: நீங்களும்தான் – நீங்களும்

Transliteration: wiingaLumthaan–wiingaLum

Translation: you are also – you too

Second is "sandhi (க்,ச்,த்,ப் - k,s,th,p) removal". Sandhi (means "joining") is a cover term for a wide variety of phonological processes that occur at morpheme or word boundaries. Examples include the fusion of sounds across word boundaries and the alteration of sounds due to neighbouring sounds or due to the grammatical function of adjacent words. Internal sandhi or morphological sandhi occurs when two morphemes are added with one another and external sandhi or syntactical sandhi occurs between two free-forms or two grammatical forms. This step removes all the external sandhi letters.

Example: அவருக்குக் - அவருக்கு

Transliteration: avarukkuk - avarukku

Translation: 'for him'

5.3. Markable Characterization

Tag analysis, morph analysis and dictionary reference are the three actions done to characterize the listed markables as listed in figure 3.

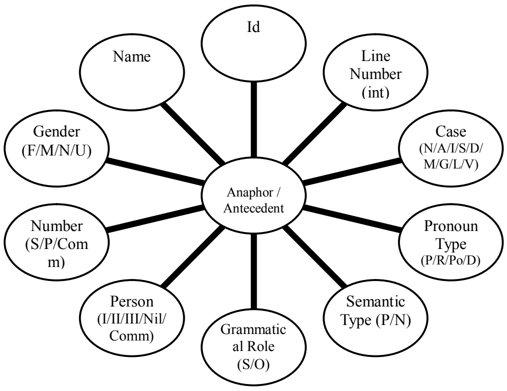

Fig. 3. Markable Characteristics

The case labels (N/A/I/S/D/M/G/L/V) and grammatical role (S-Subject / O-Object) are assigned according to Arden's Literary Tamil Case System [2] which is abbreviated in table 1. Pronoun types (P/R/Po/D) includes personal, reflexive, possessive and distributive which are already explained in chapter 3. Semantic types (P/N) of the pronoun are person and non-person. Semantic types (P/L/M/A/F/E/Pl/Li/D) of noun-phrases are person, location, material, artifact, facility, entertainment, plant, living thing and disease and are identified based on named entity classification tags. Persons of Tamil pronouns of the semantic type P are classified as I, II and III and of the semantic type N are classified as 'Nil'. There is an ambiguous situation with the reflexive pronouns [2]. Tamil number system is identified by analyzing the suffixes 'kaL' and its case morphemes. In some pronouns (Eg. அவர்கள் (avarkaL)) the plural suffix may denote honorific singulars. So the pronoun - numbers(S/P/COMM) are classified as singular, plural and common. Gender (F/M/N/U) information of both is classified as feminine, masculine, neutral and unknown. This is achieved by manually annotated look-up table (dictionary reference).

5.4. Pair Formation

Pair formation is the process of pairing each anaphor with its candidate antecedent within line distance -4 to +5, with appropriate positive or negative labelling for training. In the selected corpus, number of sentences between the anaphor and its antecedent did not exceed 5. This number may be updated accordingly, to avoid more number of negative pairs which may degrade system learning performance. Number of pairs generated in the experiment is 3288.

Table. 1. Arden's Literary Tamil Case System

Case Label	Case	Significance / grammatical role	Suffixes
N	Nominative	Subject of sentence	*[Zero]*
A	Accusative	Object of action	*-ai*
I	Instrumental	Means by which action is done	*-al*
S	Social	Association, or means by which action is done	*-out*
D	Dative	Object to whom action is performed	*(u)kku*
		Object for whom action is performed	*(u)kkaka*
M	Motion	Motion from (an inanimate object)	*-il, -ininru,-iliruntu,-iruntu*
		Motion from (an animate object)	*-iñattiliruntu*
G	Genitive	Possessive	*[Zero], -in, -utaiya,-inutaiya*
L	Locative	Place in which	*-il*
		On the person of (animate); in the presence of;	*-itam*
V	Vocative	Addressing, Calling	*e,a*

Table. 2. Feature list

Feature	Possible Values	Remarks
PAIR_ID		Unique id - generated for each pair
DIST_SENT	-4 to +5	Distance between anaphor and antecedent
ANA_CASE	N/A/I/S/D/M/G/L/V	Case marker of Anaphor
ANT_CASE	N/A/I/S/D/M/G/L/V	Case marker of antecedent
ANA-GRAMROLE	sub, obj	Grammatical role of anaphor
ANT-GRAMROLE	sub, obj	Grammatical role of antecedent
ANA_TYPE	P/R/Po/D	Type of anaphor (personal, reflexive, possessive, distributive)
PHORIC_TYPE	anaphora/cataphora	Depending on resolution direction(Backward or forward)
SEMANTICTYPE_AGREE	true, false	Depends on semantic type agreement between anaphor and antecedent
P-AGREE	true, false	Depends on person agreement between anaphor and antecedent
N-AGREE	true, false	Depends on number type agreement between anaphor and antecedent
G-AGREE	true, false, unknown	Depends on gender agreement between anaphor and antecedent
RESULT	yes, no	Depends on reference in between

5.5. Feature Vector

For each pair list of pair attributes are identified as in table 2. Example for feature vector for the pair (அவரது/PRP,தியாகராஜர்/NNP) (avarathu/PRP,thiyaagarajar/NNP) is given below:

<PAID_ID, DIST_SENT, PHORIC-TYPE, ANA_CASE, ANT_CASE, P-AGREE, N-AGREE, G-AGREE, SEMAN_TYPE_AGREE, ANA_GRAM_ROLE, ANT_GRAM_ROLE, ANA_TYPE, RESULT>
<STANJORE11, 0, ANAPH, GENI, NOMI, TRUE, TRUE, UNKNOWN, TRUE, OBJ, SUB, ANAPER-POSS, YES>

5.6. Decision Tree Classifier

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. Each interior mode corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf. A decision tree is a simple representation for classifying examples. A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same value of the target variable, or when splitting no longer adds value to the predictions.

Information gain ratio biases the decision tree against considering attributes with a large number of distinct values. So it solves the drawback of information gain—namely, information gain applied to attributes that can take on a large number of distinct values might learn the training set too well. And is calculated as follows,

Entropy at a given node t:

$$Entropy\ (t) = -\sum_{i} p\ (j \mid t) \log p\ (j \mid t \qquad (1)$$

Where $p\ (j \mid t)$ is the relative frequency of class j at node t. Entropy is maximum (log nc), when records are equally distributed among all classes implying least information and minimum (0.0) when all records belong to one class.

Information Gain:

$$Gain_{split} = Entropy\ (p) - \sum_{i=1\ to\ k} n_i/n\ Entropy\ (i) \qquad (2)$$

Where parent node, p is split into k partitions; n_i is number of records in partition i.

And Gain Ratio is,

$$GainRATIO_{split} = Gain_{split} / SplitINFO \qquad (3)$$

Where,

$$SplitINFO = -\sum_{i=1\ to\ k} n_i/n \log n_i/n \qquad (4)$$

Where parent node, p is split into k partitions, n_i is the number of records in partition i. In decision tree learning, trees are composed of decision nodes and terminal leaves. Given a new instance to be classified, test functions are applied to an instance recursively in decision nodes until hitting a leaf node which assigns a discrete output to it. A feature of the instance is tested in every node for branching. Information gain of selecting an attribute to form a tree must be calculated and a predefined number of the most informative attributes have to be selected in order to minimize the depth of the tree. In cases where more than one hypothesis is extracted from the training set, ensemble learning methods are used to increase the efficiency of the classifier by selecting and combining a set of hypotheses from the hypotheses space. These hypotheses are combined into a single classifier that makes predictions by taking a vote of its constituents. One of the mostly used methods in ensemble learning is boosting. The boosting model is sequentially induced from the training examples where the example weights are adjusted at each iteration.

6. EXPERIMENTS AND RESULTS

In this approach we applied decision tree classifier over the feature set obtained from selected data sets of tourism corpus of TDIL (NER annotated) [23].

6.1. Dataset

As there is no annotated corpus for Tamil anaphora resolution is publicly available, we have taken some portion of the tourism corpus of TDIL (Indian Language Technology Proliferation and Deployment Centre) which is annotated for Named Entity Recognition. The text contains 10,000 sentences and totally there are 2295 pronouns (excluding demonstrative pronouns) out of which the selected portion of training data contains 239 pronouns. About 22 of them are first and second person pronouns and 217 are third person pronouns. Types of pronouns involved are personal (69.8%), reflexive (7.4%) and possessive (22.8%). Though distributive pronouns do exists in Tamil text, they are rarely used in practice. This distribution is presented in figure 4.

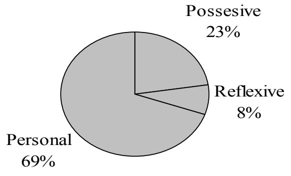

Fig. 4. Pronoun distribution in dataset

6.2. Experiments

The parameters of the decision tree classifier algorithm are set as follows:

Table 3. Sample Classification Tree

Classification Tree	Class	P(Class)	P(Target)	# Inst	Distribution (rel)	Distribution (abs)
<ROOT>	NO	0.560	0.440	678	0.560:0.440	380:298
G-AGREE = TRUE	NO	0.615	0.385	426	0.615:0.385	262:164
ANA-STYPE = ANIMATE	NO	0.622	0.378	418	0.622:0.378	260:158
PHORIC-TYPE1 = ANAPH	NO	0.601	0.399	386	0.601:0.399	232:154
ANA_GRAM_ROLE = OBJ	NO	0.643	0.357	168	0.643:0.357	108:60
ANT_GRAM_ROLE = OBJ	NO	0.619	0.381	42	0.619:0.381	26:16
ANT_GRAM_ROLE = SUB	NO	0.651	0.349	126	0.651:0.349	82:44
ANA_GRAM_ROLE = SUB	NO	0.569	0.431	218	0.569:0.431	124:94
ANT_GRAM_ROLE = OBJ	NO	0.579	0.421	38	0.579:0.421	22:16
ANT_GRAM_ROLE = SUB	NO	0.567	0.433	180	0.567:0.433	102:78
PHORIC-TYPE1 = CATAPH	NO	0.875	0.125	32	0.875:0.125	28:4
ANA-STYPE = PERSON	YES	0.750	0.750	8	0.250:0.750	2:6
ANT_GRAM_ROLE = OBJ	NO	1.000	0.000	2	1.000:0.000	2:0
ANT_GRAM_ROLE = SUB	YES	1.000	1.000	6	0.000:1.000	0:6
G-AGREE = UNKNOWN	YES	0.532	0.532	252	0.468:0.532	118:134
ANA-STYPE = ANIMATE	YES	1.000	1.000	6	0.000:1.000	0:6
ANA-STYPE = PERSON	YES	0.520	0.520	246	0.480:0.520	118:128
PHORIC-TYPE1 = ANAPH	NO	0.500	0.500	224	0.500:0.500	112:112
ANA_GRAM_ROLE = OBJ	NO	0.552	0.448	116	0.552:0.448	64:52
ANT_GRAM_ROLE = OBJ	NO	0.500	0.500	12	0.500:0.500	6:6
ANT_GRAM_ROLE = SUB	NO	0.558	0.442	104	0.558:0.442	58:46
ANA_GRAM_ROLE = SUB	YES	0.556	0.556	108	0.444:0.556	48:60
ANT_GRAM_ROLE = OBJ	YES	0.625	0.625	16	0.375:0.625	6:10
ANT_GRAM_ROLE = SUB	YES	0.543	0.543	92	0.457:0.543	42:50
PHORIC-TYPE1 = CATAPH	YES	0.727	0.727	22	0.273:0.727	6:16
ANA_GRAM_ROLE = OBJ	YES	1.000	1.000	2	0.000:1.000	0:2
ANA_GRAM_ROLE = SUB	YES	0.700	0.700	20	0.300:0.700	6:14
ANT_GRAM_ROLE = OBJ	YES	1.000	1.000	2	0.000:1.000	0:2
ANT_GRAM_ROLE = SUB	YES	0.667	0.667	18	0.333:0.667	6:12

Attribute selection: Gain Ratio
Binarization: No binarization
Pruning: 2 instances in leaves, 5 instance in node
Recursively merge leaves with same majority class: Yes
Pruning with m-estimate: m=2

The experiment is done on 3 levels of datasets (with same attribute set), DS1 with 339 instances, DS2 with 678 instances, DS3 with 1356 instances. The attribute set is as given below:

Instances: 1356
Attributes: 11 (DIST_SENT, PHORIC-TYPE, ANA_CASE, ANT_CASE, P-AGREE, N-AGREE, G-AGREE,SEMAN_TYPE_AGREE,ANA_GRAM_ROLE, ANT_GRAM_ROLE,ANA_TYPE)
Meta attributes: 1 (PAIR_ID)
Class: RESULT

In the experiment chosen validation method is 'Leave-one-out' with target class 'YES'. A sample for tree construction is expressed in table 3.

6.3. Evaluation

Table 4 shows the performance results achieved in a decision tree classifier. Below is a brief description of each of the metrics used:

Accuracy = # of correct Classifications / Total # of test instances (5)

Precision (P) =# of correct anaphoric relations produced / Total # of anaphoric relations produced (6)

Recall(R) = # of correct anaphoric relations produced / Total # of anaphoric relations in data (7)

f-Measure=(2PR)/(P+R) (8)

The accuracy metric allows for measuring the percentage of correct predictions with respect to the overall data. This metric allows one to take into accounts both positive and negative instances by paying equal attention to all types of error. The recall and precision scores indicate, respectively, errors which are caused by classifying positive instances as being negative and errors which are caused by classifying negative instances as being positive. The f-measure combines precision and recall by calculating their harmonic mean. The results according to chart in figure 5, clearly demonstrate that the learning algorithm performed much better with more training data.

Table 4. Classification Results

Dataset	No. of Instances	Parameter setting for Binarization	Classification Accuracy (CA)	F-Measure (F)	Precision (P)	Recall (R)
DS1	339	No Binarization (NB)	0.649	0.5735	0.6154	0.5369
		Exhaustive search for optimal split (Ex.Srch)	0.5664	0.4948	0.507	0.4832
		One value against others (one against)	0.6165	0.4583	0.6044	0.3691
DS2	635	No Binarization (NB)	0.7434	0.6836	0.7460	0.6309
		Exhaustive search for optimal split (Ex.Srch)	0.7493	0.7099	0.7222	0.6980
		One value against others (one against)	0.6991	0.5920	0.7327	0.4966
DS3	1356	No Binarization (NB)	0.7699	0.7329	0.7483	0.7181
		Exhaustive search for optimal split (Ex.Srch)	0.7463	0.7055	0.7203	0.6913
		One value against others (one against)	0.7139	0.6548	0.697	0.6174

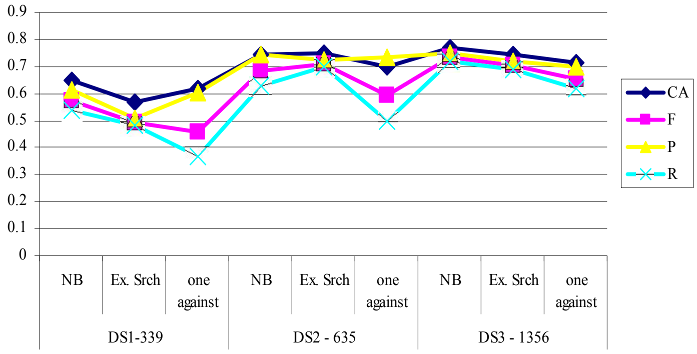

Fig. 5. Results Evaluation

7. CONCLUSION

In this paper, we present a fully fledged learning model of anaphora resolution for Tamil text, built around the decision tree model. Resolving anaphors for Tamil text is more complex than English, due to its morphological richness, semantic ambiguity and more deictic situation. We address these issues by segmentation, rule-based morphological analysis and noun phrase chunking. The mention detection and characterization components of STAR are capable of predicting candidate antecedents of personnel, reflexive, possessive and distributive pronouns and are capable of integrating a wide variety of lexical, syntactic and semantic features. Decision tree classifier is the prescribed learning model due to its simplicity and readability. The f-measure results approximately 73% with up to 77% accuracy, give the impression to be fairly satisfactory for the challenging language chore of classifying candidate antecedent of pronouns in a Tamil text. The experiments were done on NER corpus of TDIL[23] data, so as to simplify and encourage future comparisons against the results presented here. This work can be applied in almost all NLU tasks like question answering, sentiment analysis and summarization to improve their efficiency.

This work is limited only to resolve the pronouns within a document. This may further be extended to coreference chain among inter documents. That is resolving references among noun phrases of multiple documents.

REFERENCES

1. Akilandeswari A., Sobha L., "Resolution For Pronouns In Tamil Using CRF," *Proceedings Of The Workshop On Machine Translation And Parsing In Indian Languages (Mtpil-2012)*, Coling 2012, Mumbai, pp. 103–112, 2012.
2. Arden A H., *A Progressive Grammar of Common Tamil*, The Society for Promoting Christian Knowledge, Madras, 1910.
3. Balaji J., Geetha T V., Ranjani P and Madhan K., "Two-Stage Bootstrapping For Anaphora Resolution," *Proceedings of Coling 2012: Posters*, Coling 2012, Mumbai, pp. 507–516, 2012.
4. Balaji J., Geetha T V., Ranjani P. and Madhan K., "Anaphora Resolution In Tamil Using Universal Networking Language," *IICAI 2011,* pp. 1405-1415, 2011.
5. Chen Y., "Chinese Zero Anaphora Resolution and Its Applications", *Ph.D. Dissertation*, Tatung University, 2005.
6. Culotta A., Wick M. and Mccallum A., "First-Order Probabilistic Models For Coreference Resolution," *In Proceedings Of Human Language Technologies 2007: The Conference Of The North American Chapter Of The Association For Computational Linguistics*, Rochester, N.Y., 22–27 April, pp. 81–88, 2007.
7. Denis P., "New Learning Models For Robust Reference Resolution," *Doctoral Dissertation. University Of Texas at Austin,* 2007.
8. Dutta K., Prakash N. and Kaushik. S., "Machine Learning Approach for the Classification of Demonstrative Pronouns for Indirect Anaphora in Hindi News Items," *Prague Bulletin of Mathematical Linguistics.* Versita, Vol. 95, pp. 33-50. 2011.
9. Hammami S., Belguith L., and Hamadou A B., "Arabic Anaphora Resolution: Corpora Annotation with Coreferential Links," *The International Arab Journal of Information Technology(IAJIT),* Vol. 6, No. 5, pp. 481-489, 2009.
10. Hammami S., Sallemi R., Belguith L., "A Bayesian. 2010. Classifier for the Identification of Non-Referential Pronouns In Arabic," *Infos, Special Track On Natural Language Processing And Knowledge Mining,* 2010.
11. Iida R., Inui K. and Matsumoto Y., "Anaphora Resolution by Antecedent Identification Followed By Anaphoricity Determination," *ACM Trans. Asian Lang. Inf. Process*, 4, 4, 18 pp. 417-434, 2005.
12. Inoue N., Iida R., Inui K. and Matsumoto Y., "Resolving Direct And Indirect Anaphora For Japanese Definite Anaphora Pronouns," *Journal Of Natural Language Processing*.Vol. 17 No. 1, 2010.
13. K11lcaslan Y., Guner E S. and Y1ld1r1m S., "Learning-Based Pronoun Resolution for Turkish with a Comparative Evaluation," *Computer Speech and Language*. No. 23, pp. 311–331, 2009.
14. Mccarthy J F., and Lehnert W G., "Using Decision Trees For Coreference Resolution," *In Proceedings Of The 14th International Joint Conference On Artificial Intelligence (IJCAI)*, pp. 1050–1055, 1995.
15. Narayana Murthi K., Sobha L. and Muthukumari B., "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, UK, pp. 39-50, 2007.
16. Pal T L., Dutta K. and Singh P., "Anaphora Resolution In Hindi: Issues and Challenges," *International Journal of Computer Applications.* (0975 – 8887) Vol. 42, No.18, 2012.
17. Sasano R., "Japanese Anaphora Resolution Based On Automatically Acquired World Knowledge," *Doctoral Dissertation*. 2008.
18. Sobha L., "Anaphora Resolution In Malayalam And Hindi," *Unpublished Doctoral Dissertation. Mahatma Gandhi University*, Kottayam, Kerala, 1999.
19. Soon W M., Ng H T., and Lim D C Y., "A Machine Learning Approach to Coreference Resolution Of Noun Phrases," *Computational Linguistics.* Vol. 27, No. 4, pp. 521–544, 2001.
20. Yang X., Su J., Lang J., Tan C., Liu T., and Li S., "An Entity-Mention Model For Coreference Resolution With Inductive Logic Programming," In *Proceedings Of Acl-08: Hlt. Columbus (Oh). Association for Computational Linguistics*. pp. 843–51, 2008.
21. Zitouni I., Luo X. and Florian R., "A Cascaded Approach To Mention Detection And Chaining In Arabic," *IEEE Transactions On Audio, Speech And Language Processing,* Vol. 17, No. 5, 2009.
22. Zitouni, I. and Florian, R., "Cross Language Information Propagation for Arabic Mention Detection," *ACM Trans. Asian Lang. Inform. Process, 8, 4, Article 17 (December 2009), 2009.*
23. TDIL (2013) Linguistic Resources namely Tourism Text Corpora -Tamil Department of Electronics and Information Technology (DeitY), Government of India, Language Technology Proliferation and Deployment Centre, New Delhi http://tdil-dc.in//tdildcMain/material/765104hin_tam_tourism_set09.zip (accessed 23 May 2013).