Resolving Pronouns for a Resource-Poor Language, Malayalam Using Resource-Rich Language, Tamil

Sobha Lalitha Devi
AU-KBC Research Centre, Anna University, Chennai
sobha@au-kbc.org

Abstract

In this paper we give in detail how a resource rich language can be used for resolving pronouns for a less resource language. The source language, which is resource rich language in this study, is Tamil and the resource poor language is Malayalam, both belonging to the same language family, Dravidian. The Pronominal resolution developed for Tamil uses CRFs. Our approach is to leverage the Tamil language model to test Malayalam data and the processing required for Malayalam data is detailed. The similarity at the syntactic level between the languages is exploited in identifying the features for developing the Tamil language model. The word form or the lexical item is not considered as a feature for training the CRFs. Evaluation on Malayalam Wikipedia data shows that our approach is correct and the results, though not as good as Tamil, but comparable.

1 Introduction

Natural language processing techniques of the present day require large amounts of manually-annotated data to work. In reality, the required quantity of data is available only for a few languages of major interest. In this work we show how a resource-rich language, Tamil, can be leveraged to resolve anaphora for a related resource-poor language, Malayalam. Both Tamil and Malayalam belong to the same language family, Dravidian. The methods we focus on exploits the similarity at the syntactic level of the languages and anaphora resolution heavily depends on syntactic features. If the resources available in one language (henceforth referred to as source) can be used to facilitate the resolution, such as anaphora, for all the languages related to the language in question (target), the problem of unavailability of resources would be alleviated.

There exists a recent research paradigm, in which the researchers work on algorithms that can rapidly develop machine translation and other tools for an obscure language. This work falls into this paradigm, under the assumption that the language in question has a less obscure sibling. Moreover, the problem is intellectually interesting. While there has been significant research in using resources from another language to build, for example, parsers, there have been very little work on utilizing the close relationship between the languages to produce high quality tools such as anaphora resolution (Nakov, P and Tou Ng,H 2012). In our work we are interested in the following questions:

If two languages are from the same language family and have similarity in syntactic structures and not in lexical form and script

1. Can the language model developed for one language be used for analyzing the other language?
2. How the lexical form difference can be resolved in using the language model?
3. How to overcome the challenges of script variation?

In this work we develop a language model for resolving pronominals in Tamil using CRFs and using the language model test another language, Malayalam. As said earlier, in this work, we are

Proceedings of Recent Advances in Natural Language Processing, pages 611–618,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_072

focusing on Dravidian family of languages. Historically, all Dravidian languages have branched out from a common root Proto-Dravidian. Among the Dravidian languages, Tamil is the oldest language. Though there is similarity at the syntactic level, there is no similarity in lexical form or at the script level among the languages. We are motivated by the observation that related languages tend to have similar word order and syntax, but they do not have similar script or orthography. Hence words are not similar.

Tamil has the most resources at all levels of linguistics, right from morphological analyser to discourse parser and Malayalam has the least. About the similarity of the two languages we give in detail in section 2.

The remainder of this article is organized as follows: Section 2 provides a detailed description of the two languages, their linguistic similarity credits and differences, Section 3 presents the pronominal resolution in Tamil. In Section 4 we introduce our proposed approach on how the language model for Tamil can be used for resolving Malayalam pronouns. Section 5 describes the datasets used for evaluation, experiments and analysis of the results and the paper ends with the conclusion in Section 6.

2 How similar the languages Tamil and Malayalam Are?

As mentioned earlier, both the languages belong to the Dravidian family and are relatively free word order languages, inflectional (rich in morphology), agglutinative and verb final. They have Nominative and Dative subjects. The pronouns have inherent gender marking as in English and have the same lexical form "avan" "he", "aval" "she" and "atu" "it" both in Tamil and Malayalam. Though the pronouns have same lexical form and meaning, it can be said that there is no lexical similarity between the two languages. The similarity between two languages can be at three levels, a) writing script, b) the word forms and c) the syntactic structure.

Script Level: The two languages have different writing form, though the base is from Grandha script. Hence no similarity at the script level.
The Word Level: There is no similarity at the lexical level between the two languages. The Sanskritization of Malayalam contributed to have more Sanskrit verbs in Malayalam whereas Tamil retained the Proto –Dravidian verbs. For example, the word for "talk" in Malayalam is "samsarikkuka" "to talk" which has root in Sanskrit, whereas the Tamil equivalent is "pesuu" "to talk", the root in Pro-Dravidian.
The Syntactic Structure Level: There is lot of similarity at the syntactic structure level between the two languages. Since antecedent to anaphor has dependency on the position of the noun, the structural similarity is a positive feature for our goal. The syntactic similarity at Sentence level, Case maker level, pronominal distribution level are explained with examples.
Case marker level: Both the languages have the same number of cases and their distribution is similar. In both the languages, nouns inflected with nominative or dative case become the subject of the sentence (Dative subject is the peculiarity of Indian languages). Accusative case denotes the object.
Clausal sentences: The clause constructions in both the languages follow the same rule. The clauses are formed by nonfinite verbs. The clauses do not have free word order and they have fixed positions. Order of embedding of the subordinate clause is same in both the languages.

Ex:1

(Ma) [innale vanna(vbp) kutti]/Sub-RP-cl
{sita annu}/Main cl

(Ta) [neRu vanta(vbp) pon]/Sub- RP-cl
{sita aakum}/Maincl

[Yesterday came girl]/subcl
{Sita is}/ Maincl

(The girl who came yesterday is Sita)

As can be seen from the above example, the basic syntactic structure is the same in both the languages. The above example is a two clause sentence with a relative participial clause and a main clause. The relative participial clause is formed by the nonfinite verb (vbp). Using the same example we can find the pronominal distribution.

Ex: 2

(Ma) [innale vanna(vbp) $aval_i$ (PRP)
]/Sub-RP-cl {$sita_i$ annu}/Main cl
(Ta) [neRu vanta(vbp) $aval_i$ (PRP)
]/Sub-RP-cl {$sita_i$ aakum}/Maincl
[Yesterday came she_i (PRP)
]/subcl { $Sita_i$ is} / Maincl
(The she who came yesterday is Sita)

In the above example the pronoun "aval" "she" occurs at the same position in both the languages

and the antecedent "sita" also occurs at the same position as shown by co-indexing. Consider another example.

Ex:3

(Ma). sithaa$_i$ kadaikku pooyi. aval$_i$ pazham Vaangicchu(Vpast)

(Ta). sithaa$_i$ kadaikku cenRaal. aval$_i$ pazham Vaangkinaal(V,past,+F,+Sg).

Sita shop went. She fruit bought

(Sita$_i$ went to the shop. She$_i$ bought fruit.)

In the above example there are two sentences and pronoun is in one sentence and antecedent is in another. Here you can see the distribution of the pronoun "aval" and where the antecedent "sita" is occurring. Though Tamil has number, gender and person agreement between subject and verb and Malayalam does not have, this cannot be considered as a grammatical feature which can be used for identifying the antecedent of an anaphor. This grammatical variation does not have an impact on the identification of pronoun and antecedent relations. From the above examples we can see that the two languages have the same syntactic structure at the clause and sentence level. We are exploiting this similarity between the two languages to achieve our goal. We find that using this similarity between the languages, the language model of Tamil can be used to resolve pronouns in Malayalam.

3 Pronoun Resolution in Tamil

3.1 Pronouns in Tamil

In this section, we analyse in detail the pronominal expressions in Tamil. Pronouns are the words used as a substitution to nouns, that are already mentioned or that is already known. There are pronouns which do not refer. Pronouns in Tamil have person (1st, 2nd, 3rd person) and number (singular, plural) distinction. Masculine, feminine and neuter gender distinctions are clearly marked in 3rd person pronouns, whereas in 1st and 2nd person pronouns there is no distinction of masculine, feminine and neuter gender. In this work we consider only third person pronouns. Third person pronouns in Tamil have inherent gender and as in English and they are "avan" he, "aval" she and "atu" it. In this work, we resolve 3rd person pronouns. The distribution of pronouns in various syntactic constructions is explained with examples below.

Ex:4

4a. ***maaNavarkaL***$_i$ *paLLikku celkiranar.*

Students(N) school(N)+dat go(V)+present+3pl

(Students are going to the school)

4b. ***avarkaL***$_i$ *veekamaaka natakkinranar.*

They(PN) fast(ADV) walk(V)+present+3pl

(They are walking fast.)

Considering Ex 4a and Ex.4b, sentence Ex.4b has 3rd person plural pronoun 'avarkaL' as the subject and it refers to plural noun 'maaNavarkaL' which is the subject in Ex.4a.

Ex:5

5a. ***raamuvum***$_i$ ***giitavum***$_j$ *nanparkaL.*

Raamu(N)+INC Gita(N)+INC friends(N)

(Ramu and Gita are friends.)

5b. ***avan***$_i$ *ettaam vakuppil padikkiraan.*

He(PN) eight(N) class(N) study(V)+present+3sm

(He studies in eight standard.)

5c. ***avaLum***$_j$ *ettaam vakuppil padikkiaal.*

She(PN) eight(N) class(N) study(V)+present+3sf

(She also studies in eight standard.)

In Ex.5b, 3rd person masculine pronoun 'avan' occurs as the subject and it refers to the masculine noun 'raamu', subject noun in Ex.5a. Similarly, 3rd person feminine pronoun 'avaL' in Ex.5c refers to feminine noun 'giita' in Ex.3a. 'atu', which is a 3rd person neuter pronoun, will also occurs as genitive/possessive case marker. Consider the following example.

3.1.1 Non-anaphoric Pronouns

The pronouns can also occur as generic mentions without having referent. In English it known as'pleonastic it'.

Ex:6.

atu *oru malaikalam.*

It(PN) one(Qc) rainy_season (N)

(It was a rainy season.)

In Ex.6, the 3rd person neuter pronoun 'atu' (it) do not have a referent. Here 'atu' is equivalent to pleonastic 'it' in English

3.1.2 Corpus annotation

We collected 600 News articles from various online Tamil News wires. The News articles are from Sports, Disaster and General News domains. The anaphoric expressions are annotated along with its antecedents using graphical tool, PAlinkA, a highly customisable tool for Discourse Annotation (Orasan, 2003) which we customized for Tamil. We have used two tags namely, MARKABLE and COREF. The corpus used for training is 54,563 words which are annotated for anaphora –antecedent pairs and testing

corpus is 10,912 words. We have calculated the inter-annotator agreement which is the degree of agreement among annotators. We have used 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. We got a Kappa score of 0.87. The difference between the annotators 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.

3.2 Pronoun Resolution System

Early works in anaphora resolution by Hobbs (1978), Carbonell and Brown (1988), Rich and Luper Foy (1988) etc. were mentioned as knowledge intensive approach, where syntactic, semantic information, world knowledge and case frames were used. Centering theory, a discourse based approach for anaphora resolution was presented by Grosz (1977), Joshi and Kuhn (1979). Salience feature based approaches were presented by Lappin and Leass (1994), Kennedy Boguraev (1996) and Sobha et al., (2000). Indicator based, knowledge poor method for anaphora resolution methods were presented by Mitkov (1997, 1998). One of the early works using machine learning technique was Dagan Itai's (1990) unsupervised approach based on co-occurrence words. With the use of machine learning techniques researchers work on anaphora resolution and noun phrase anaphora resolution simultaneously. The other machine learning approaches for anaphora resolution were the following. Aone and Bennett (1995), McCarty and Lahnert (1995), Soon et al., (2001), Ng and Cardia (2002) had used decision tree based classifier. Anaphora resolution using CRFs was presented by McCallum and Wellner (2003) for English, Li et al., (2008) for Chinese and Sobha et al., (2011, 2013) for English and Tamil.

In Indian languages anaphora resolution engines are demonstrated only in few languages such as Hindi, Bengali, Tamil, and Malayalam. Most of the Indian languages do not have parser and other sophisticated pre-processing tools. The earliest work in Indian language, 'Vasisth' was a rule based multilingual anaphora resolution platform by Sobha and Patnaik (1998, 2000, 2002), where the morphological richness of Malayalam and Hindi were exploited without using full-parser. The case marker information is used for identifying subject, object, direct and in-direct object. Prasad and Strube (2000), Uppalapu et al., (2009) and Dekwale et al., (2013) had presented different approaches using Centering theory for Hindi. Sobha et al., (2007) presented a salience factor based with limited shallow parsing of text. Akilandeswari et al., (2013) used CRFs for resolution of third person pronoun. Ram et al., (2013) used Tree CRFs for anaphora resolution for Tamil with features from dependency parsed text. In most of the published works resolution of third person pronoun was considered and it is a non-trivial task.

Pronoun resolution engine does the task of identifying the antecedents of the pronouns. The Pronoun resolution is built using Conditional Random Fields (CRFs) technique. Though CRFs is notable for sequence labelling task, we used this technique to classify the correct anaphor-antecedent pair from the possible candidate NP pairs by presenting the features of the NP pair and by avoiding the transition probability. While training we form positive pairs by pairing anaphoric pronoun and correct antecedent NP and negative pairs by pairing anaphoric pronouns and other NPs which match in person, number and gender (PNG) information with the anaphoric pronoun. These positive and negative pairs are fed to the CRFs engine and the language model is generated. While testing, when an anaphoric pronoun occurs in the sentence, the noun phrases which match in PNG with the pronoun, that occur in the preceding portion of the sentence and the four preceding sentences are collected and paired with the anaphoric pronoun and presented to CRFs engine to identify the correct anaphor-antecedent pair.

3.2.1 Pre-processing

The input document is processed with a sentence splitter and tokeniser to split the document into sentences and the sentences into individual tokens which include words, punctuation markers and symbols. The sentence split and tokenized documents are processed with Syntactic Processing modules. Syntactic processing modules include Morphological analyser, Part-of-Speech tagger and Chunker. These modules are developed in house.

a) **Morphological Analyser:** Morphological analysis processes the word into component morphemes and assigning the correct morpho-syntactic information. The Tamil morphological Analyser is developed using paradigm based approach and implemented using Finite State Automata (FSA) (Vijay Sundar et.al 2010). The words are classified according to their suffix formation and are marked as paradigms. The number of paradigms used in this system is Noun

paradigms: 32; Verb Paradigms: 37; Adjective Paradigms: 4; Adverb Paradigms: 1. A root word dictionary with 1, 52,590 root words is used for developing the morphological anlyser. The morphological analyser is tested with 12,923 words and the system processed 12,596 words out of which it correctly tagged 12,305. The Precision is 97.69% and Recall = 97.46%. MA returns all possible parse for a given word.

b) **Part Of Speech Tagger (**POS tagger): Part of Speech tagger disambiguates the multiple parse given by the morphological analyser, using the context in which a word occurs. The Part of speech tagger is developed using the machine learning technique Conditional Random fields (CRF++) (Sobha L, et.al 2010). The features used for machine learning is a set of linguistic suffix features along with statistical suffixes and uses a window of 3 words. We have used 4, 50,000 words, which are tagged using BIS POS tags. The system performs with recall 100% and Average Precision of 95.16%.

c) **Noun and Verb Phrase Chunker:** Chunking is the task of grouping grammatically related words into chunks such as noun phrase, verb phrase, adjectival phrase etc. The system is developed using the machine learning technique, Conditional Random fields(CRF++) (Sobha L et.al 2010). The features used are the POS tag, Word and window of 5 words. Training Corpus is 74,000 words. The recall is 100%. Average Precision of 92.00%.

3.3 Pronoun Resolution Engine

In both training and testing phase, the noun phrases (NP) which match with the PNG of the pronoun are considered. The features are extracted from these NPs. In the training phase the positive and negative pairs are marked and fed to the ML engine for generating a language model. In the testing phase these NPs with its features are input to the language model to identify the antecedent of a pronoun. Here we have not taken the lexical item or the word as a feature. We have used only the grammatical tags as feature. The features selected represent the syntactic position of the anaphor –antecedent occurrence.

3.3.1 Features Selection

The features required for machine learning are identified from shallow parsed input sentences. The features for all possible candidate antecedent and pronoun pairs are obtained by pre-processing the input sentences with morphological analyser, POS tagger, and chunker. The features identified can be classified as positional and syntactic features

Positional Features: The occurrence of the candidate antecedent is noted in the same sentence where the pronoun occurs or in the prior sentences or in prior four sentences from the current sentence.

Syntactic Features: The syntactic arguments of the candidate noun phrases in the sentence are a key feature. The arguments of the noun phrases such as subject, object, indirect object, are obtained from the case suffix affixed with the noun phrase. As mentioned in section 2 the subject of a sentence can be identified by the case marker it takes. We use morphological marking for the above.

a) PoS tag and chunk tag of Candidate NP, case marker tags of the noun.
b) The suffixes which show the gender which gets attached to the verb.

3.3.2 Development of Tamil Language Model

We used 600 Tamil Newspaper articles for building the language model. The preparation of the training data is described below. The raw corpus is processed with sentence splitter and tokeniser. The tokenized corpus is then preprocessed with shallow processing modules, namely, morphological analyser, part-of-speech tagger and chunker. The training data is prepared from this processed corpus. For each pronoun, the Noun phrase (NP) preceding the pronouns and in the NPs in preceding sentence till correct antecedent NP, which match in Person, number and Gender (PNG) are selected for training. The above features are used for CRF for learning. The system was evaluated with data from the web and the result is given below.

Domain	Testing Corpus (Words)	Precision (%)	Recall (%)
News data	10,912	86.2	66.67

Table 1: Pronominal Resolution (CRFs engine)

4. Resolution of Pronouns in Malayalam Corpus using Tamil Language Model

In this section, we present in detail how Malayalam is tested using Tamil language model. Here Malayalam data is pre-processed as per the requirement of the Tamil language model test data. The test data required four grammatical information, i) POS, ii) the case marker, iii) the number gender and person and iv)chunk information. In the introduction we have asked three questions on how to use a language model in source language be used for testing a target language. The three questions are dealt one by one below.

1. Can the language model developed for one language be used for analyzing the other language?

In this study we have used the language model of the source language Tamil. The features used to develop this language model are POS tag, Chunk tag and the case/ suffix tags. The word form was not considered as a feature. Since these are the features used for learning, the test data also should have these information. As said earlier, Malayalam is not a resource rich language and it does not have pre-processing engines such as POS tagger, Chunker and morphological analysers with high accuracy. Hence we developed a very rudimentary preprocessing systems which can give the POS tag, case/suffix tags and chunk tags.

The POS information: The POS and suffix information are assigned to the corpus using a root word dictionary with part of speech information and a suffix dictionary.

The dictionary has the root words which include all types of pronouns and contains nearly 66,000 root words. The grammatical information (POS) such as noun, verb, adjective, pronoun and number gender person (PNG) information are given for a word in the dictionary.

The suffix dictionary contains all possible suffixes which a root word can take. The suffix includes the changes in sandhi when added to the root word. The suffixes are of two types, i) that which gets attached to nouns called the case suffixes and 2) that which gets attached to verbs called the TAM (Tense, Aspect, and Modal). Using this suffix dictionary we can identify the POS of the word even if the word is not present in the root word dictionary. The suffix dictionary has 1,00,000 unique entries.

The noun chunk information is given by a rule based chunker which works on three linguistic rules. The noun phrases alone are required for anaphora resolution. Chunks are identified using the basic linguistic rule for Noun phrases as given below

1. [determiner] [quantifier][intensifier] [classifier][adjective] {Head Noun}. Here Head noun is obligatory and others are all optional
2. NN+NN combination
3. NN with no suffix+NN with no suffix…… +NN with suffix or without suffix.

Using the above rules we identified the noun chunks in Malayalam. The above discussed preprocessing gives an accuracy of 66% for POS and suffix tagging and 63% for chunking.

2. How the lexical form difference can be resolved in using the language model?

The second question we asked is about the lexical form or words which are not similar in both the languages and how this can be resolved. The analysis of Tamil has shown that the syntactic structure of the language has more prominence over the words in the resolution of anaphors. Hence we have taken the syntactic features and did not take word as a feature. The system learned only the structure patterns.

3. How to overcome the challenges of script variation?

Since word feature is not considered the script do not pose any challenges. Still to have the same script we converted the two languages into one form the WX notation. This helped in having the same representation of the languages.

4. Testing with Malayalam Data

We selected 300 articles from Malayalam Wikipedia, which were on different genre and size. The 300 Malayalam documents from Wikipedia has 7600 sentences with 3660 3rd person pronouns. The distribution of the pronouns is presented in Table 2.

Pronoun	Number of Occurrences with its Inflected forms
avan (3rd person masculine singular)	1120
aval (3rd person feminine singular)	840
avar (3rd person honorific)	420
athu (3rd person nuetor singular)	1280
Total	3660

Table 2. Distribution of 3rd person pronouns in the corpus

As discussed in the earlier section, the preprocessing done for Tamil using syntactic module are morphological information, POS and Chunking information. Hence the same pre-processing information is necessary for Malayalam data as

well. The documents are initially preprocessed with a sentence splitter and tokenizer. The tokenized documents are pre-processed using the pos, suffix and chunking systems discussed above to enrich the text with syntactic information.
For each pronoun, we identify the possible candidate antecedents. Those noun phrases that occur preceding the pronoun in the current sentences and preceding four sentences, which match in the person, number gender (PNG) with the select pronoun are identified as possible candidates. For these possible candidates we extract the features required for CRFs techniques as explained in the previous section. After extraction of features for selected candidate antecedents, the antecedent is identified by using language model built using Tamil data. The results are encouraging with 67% accuracy which is a respectable score. The errors and evaluation is given in detail in the next section.

5. Experiment and Discussion

The experiment showed that the resolution of the pronouns "avan" he and "aval" she is similar to that of Tamil documents. The issues related to split antecedent is not addressed and hence pronouns which are referring to coordinated nouns were not resolved. The pronoun which is less resolved is the third person neuter pronouns "atu" compared to other pronouns. The third person neuter pronoun usually has more number of possible candidates, which leads to poor resolution. Consider the following example.

Ex:7

avan joli ceytha jolikkarkku
He(PRP) work(N) do(V)+past+RP worker(N)+pl+dat
vellam kotuthu.
water(N) give(V)+past
(He gave water to the workers who did the work.)

atu nallatu aayirunnu
It(PRP) good(N) is (Copula V) +past.
(It was good.)

In this example, the pronoun 'atu' refers to 'vellam" 'water'; in the previous sentence. In the previous sentence there are two possible candidate antecedents 'vellam' and 'joli 'which are 3rd person nouns. The ML engine chooses 'joli', which is in the initial position of the sentence and in the subject position. When the antecedent is in the object position the engine has not identified it properly. The following table gives the results of pronouns.

Type of pronoun	Precision (%)	Recall (%)
avan (3rd person masculine singular)	70.83	69.52
avaL (3rd person feminine singular)	69.34	68.56
avar/avarkal (3rd person plural/honorofic)	65.45	70.34
atu (3rd person nuetor singular)	56.67	65.67
Total	68.45	67.34

Table 3: Evaluation Results

6. Conclusion

In this paper we explained a method to use high resource language to resolve anaphors in less resource language. In this experiment the high resource language is Tamil and the less resource language is Malayalam. The results are encouraging. The model needs to be tested with more data as future work.

References

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

Dagan I., and Itai. A. 1990. *Automatic processing of large corpora for the resolution of anaphora references*. In: 13th conference on Computational linguistics, Vol. 3, Helsinki, Finland, pp.330-332.

Dakwale. P., Mujadia. V., Sharma. D.M. 2013. *A Hybrid Approach for Anaphora Resolution in Hindi.* In: Proc of International Joint Conference on Natural Language Processing, Nagoya, Japan, pp.977–981.

Li., F., Shi., S., Chen., Y., and Lv, X. 2008. *Chinese Pronominal Anaphora Resolution Based on Conditional Random Fields.* In: International Conference on Computer Science and Software Engineering, Washington, DC, USA, pp. 731-734.

Hobbs J. 1978. *Resolving pronoun references*. Lingua 44, pp. 339-352.

Grosz, B. J. 1977. *The representation and use of focus in dialogue understanding. Technical Report* 151, SRI International, 333 Ravenswood Ave, Menlo Park, Ca. 94025.

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.

Kennedy, C., Boguraev, B. 1996 *Anaphora for Everyone: Pronominal Anaphora Resolution without a*

Parser. In: 16th International Conference on Computational Linguistics COLING'96, Copenhagen, Denmark, pp. 113–118.

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

McCallum A., and Wellner. B. 2003. *Toward conditional models of identity uncertainty with application to proper noun coreference*. In Proceedings of the IJCAI Workshop on Information Integration on the Web, pp. 79–84.

McCarthy, J. F. and Lehnert, W. G. 1995. *Using decision trees for coreference resolution*. In C. Mellish (Ed.), Fourteenth International Conference on Artificial Intelligence, pp. 1050-1055

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.

Mitkov, R. 1997. "*Factors in anaphora resolution: they are not the only things that matter. A case study based on two different approaches*". In Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution, Madrid, Spain.

Ng V., and Cardie C. 2002. *Improving machine learning approaches to coreference resolution*. In. 40th Annual Meeting of the Association for Computational Linguistics, pp. 104-111.

Prasad R., and Strube,M.,2000. *Discourse Salience and Pronoun Resolution in Hindi*, Penn Working Papers in Linguistics, Vol 6.3, pp. 189-208.

Orasan, C. 2003*, PALinkA: A highly customisable tool for discourse annotation*. SIGDIAL Workshop 2003: 39-43

Preslav Nakov Hwee Tou Ng 2012. *Improving Statistical Machine Translation for a Resource-Poor Language Using Related Resource-Rich Languages*; Journal of Artificial Intelligence Research 44 (2012) 179-222;

Ram, R.V.S. and Sobha Lalitha Devi. 2013.*"Pronominal Resolution in Tamil Using Tree CRFs*", In Proceedings of 6th Language and Technology Conference, Human Language Technologies as a challenge for Computer Science and Linguistics - 2013, Poznan, Poland

Rich, E. and LuperFoy S.,1988 *An architecture for anaphora resolution*. In: Proceedings of the Second Conference on Applied Natural Language Processing, Austin, Texas.

Russell, B. 1919, *"On Propositions: What They Are and How They Mean,"* Proceedings of the Aristotelian Society, Supplementary Volume 2: 1–43; also appearing in Collected Papers, Vol. 8

Senapati A., Garain U. 2013. *GuiTAR-based Pronominal Anaphora Resolution in Bengal*. In: Proceedings of 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria pp 126–130.

Sikdar U.K, Ekbal A., Saha S., Uryupina O., Poesio M. 2013. *Adapting a State-of-the-art Anaphora Resolution System for Resource-poor Language*. In proceedings of International Joint Conference on Natural Language Processing, Nagoya, Japan pp 815–821.

Sobha L. and Patnaik B. N. 2000. *Vasisth: An Anaphora Resolution System for Indian Languages*. In Proceedings of International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications, Monastir, Tunisia.

Sobha L. and Patnaik,B.N. 2002. *Vasisth: An anaphora resolution system for Malayalam and Hindi*. In Proceedings of Symposium on Translation Support Systems.

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

Sobha L., Sivaji Bandyopadhyay, Vijay Sundar Ram R., and Akilandeswari A. 2011. *NLP Tool Contest @ICON2011 on Anaphora Resolution in Indian Languages*. In: Proceedings of ICON 2011.

Sobha Lalitha Devi and Pattabhi R K Rao. 2010. *Hybrid Approach for POS Tagging for Relatively Free Word Order Languages.* In Proceedings of Knowledge Sharing Event on Part-Of-Speech Tagging, CIIL, Mysore.

Vijay Sundar Ram and Sobha Lalitha Devi. 2010. *Noun Phrase Chunker Using Finite State Automata for an Agglutinative Language.* In Proceedings of the Tamil Internet – 2010, Coimbatore, India, 218–224.

Soon W. H., Ng, and Lim D. 2001. *A machine learning approach to coreference resolution of noun phrases.* Computational Linguistics 27 (4), pp.521-544.

Taku Kudo. 2005. CRF++, an open source toolkit for CRF, http://crfpp.sourceforge.net .

Uppalapu. B., and Sharma, D.M. 2009. *Pronoun Resolution For Hindi*. In: Proceedings of 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 09), pp. 123-134.