Anaphora Resolution from Social Media Text in Indian Languages (SocAnaRes-IL) : 2nd Edition-Overview

Sobha Lalitha Devi[a]

[a] *AU-KBC Research Centre, MIT Campus of Anna University, Chennai, India*

Abstract

Anaphora and its antecedent are to be identified for Natural Language Understanding (NLU) applications such as Information Extraction, Conversation Analysis, Opinion Mining, Machine Translation etc. There is a great need to develop applications such as Anaphora resolution, co-reference resolution which can be used in NLU systems. This Second Edition of shared task on Anaphora resolution from the microblog text and conversation for languages such as Hindi, Tamil and Malayalam (Indian Languages) is similar to first edition with more microblog data and data on conversation. The aim is to provide annotated data in anaphora and enhance the research in this area. In this edition too we gave data from Hindi, Tamil, Malayalam from Indian languages and also from English which can be used as resource rich language, if one wants to take Indian languages as resources poor language.

There were six registered groups who took data for development and testing but only one group submitted the run. They have used Neutralcoref network by huggingface.

Keywords 1

Anaphora Resolution, Social Media Text Analysis, Indian Languages, Hindi, Malayalam, Tamil, English, Machine Learning

1. Introduction

The second edition of SocAnaRes-IL is similar to its first edition in all its objectives. The difference is that more data in microblog is provided and also included conversation data which are manually created in Malayalam and translated to other languages.

The social media platforms such as Twitter has generated large amount of microblog texts bringing in a new challenge for NLP applications, which asks for a new perspective in language technology research. These microbolg texts present a discourse genre which carries non-standard language characteristics such as noisy or informal language, abbreviations which do not follow the regular patterns, purposeful typos or spellings, use of non-alphanumerical symbols such as # and @ etc. This requires new methodologies and techniques for processing such texts. The challenges brought in by these texts are spread on all aspects of language computing right from developing the tag, collecting the corpora, annotation of the corpus and the development of the system. The task proposed here is to develop an anaphora resolution (AR) system from the Twitter data annotated for anaphor and its antecedent. The languages considered are from Indo Aryan and Dravidian families and they are Hindi, Tamil and Malayalam respectively and English. The objectives of the task are:

- Creation of benchmark data for Anaphora Resolution in Indian language microblog texts -Twitter data.
- Encourage researchers to develop novel systems for Anaphora Resolution.

FIRE 2022: Forum for Information Retrieval Evaluation, December 9-13, 2022, Kolkata, India

EMAIL: sobha@au-kbc.org
ORCID: 0000-0003-1326-9474

- Providing opportunity to researchers to have comparison of different techniques.

Anaphora resolution has been a challenging area in research and has been going on for more than 4 decades and the challenges became more when it shifted from normal texts to microblog texts. There is very little work done for Indian languages. Most prominent Indian languages which have good anaphora resolution systems are Malayalam, Tamil, Bengali and Hindi. As in every conversation, anaphors are extensively used in microblog texts as well, only with the difference that their usage differs from that in normal text. To give an example, the antecedent falling outside the text is a very common occurrence. Similarly the antecedent not being a noun phrase but a hash tag or to an earlier tweet is very common. It can also refer to an event that is being trolled which need not be explicitly marked in the current tweet. Among the types of anaphors, pronominals are widely used.

The approaches used for anaphora resolution are many and they can be classified as rule based approach, Corpus based approach using Machine learning technique, Knowledge poor approach and discourse based approach. The knowledge poor approach of Mitkov [4], Approach with no deep parsing of Kenndey & Baguraev [3] and salience based approach of Lappin & Leas [5] are the most prominent among them. The recent works use machine learning techniques such as decision tree, CRFs [1-2] and Tree CRFs [8]. There are some works done where resource rich language is used for resolving pronominals in less resource language [7].

In this task, the data was collected from Twitter using Twitter API and annotated for Anaphor and Antecedent pair alone. The method of collecting data is same as in the first edition of SocAnaRes2020. In this edition we have included conversation data. The conversation data was created manually for Malayalam and translated to other languages. The annotation details are given in section on 2, the corpus does not have any other grammatical annotation such as POS or NP/VP Chunk. The participants were free to use any external resources and any method.

2. Corpus Compilation

There are various challenges in anaphora resolution on microblog texts from Twitter. The language of Twitter language belongs to the Computer Mediated Communication (CMC) or Technology Mediated Communication (TMC) languages, where there are restrictions in rendering. Tweets have fixed character limitations and users are forced to communicate their idea using this limited number of character. Hence there is language variation and various types of word and syntactic level variations are brought in to accommodate the idea with in the given character span. And thus the language we are analyzing are different in morphology and syntax such as non-standard language characteristics such as noisy or informal language, abbreviations which do not follow the regular patterns, purposeful typos , New spelling, use of non-alphanumerical symbols such as # and @ , use of symbols such as emoticons/emojis , use of meta tags and hash tags . Another import aspect of twitter data is Code mixing. The code mix is at the word level and also at the script level. Yet another characteristic is the dialectal variations, which are inherent to all languages, are also seen in twitter data. The dialects can be of various types such as regional, religious and community based. The users tend to use their dialect and the words they use may not be there in a dictionary. We need to preprocess the data to normalize the vocabulary.

Tweets are generally very short and lack sufficient context to determine an antecedent of an anaphor. Especially in the resolution of third person pronominals "he/they" (woh, ve, vo; avar), in atleast 20% of the cases, the antecedent is not mentioned in the current Tweet, it is either in posts which was already said a day before or is present in the troll. And is understood with world knowledge. An example Tweet is given below:

HI: "@vijayrk modi sarkar ke baad garibi kam hui hai, bank wale ab usko bhi loan dena shuru kiya hai"

("@vijayrk after Modi government poverty has reduced, now banks are giving loans to them")

Here in this tweet “usko” is the third person pronoun, and here it referring to poor people. The antecedent for this pronoun can be identified only if we have world knowledge.

2.1. Corpus Collection

The corpus was collected using Twitter API. Our aim was to collect conversations by recursively retrieving the parent tweets to construct the full conversational tree structure. The tweets for training were collected on 2 days in the month of June and for testing collected on 2 days in the month of August for each language. First a set of tweets were collected using the event key phrases of that day such as “election campaign of US”, “Government announcements” in the respective languages. After the first set of tweets are collected, we identify if these tweets are retweets or reply tweets to any other tweet, using “reply_ to _id” field of the Tweet data structure (or called as Tweet Object) of the Twitter API. In this work we have not taken the re-tweets. The tweets which are reply tweets (let us call this as RPT), for those we identify the original tweet (let us call this as ORT) for which this tweet is given as reply, using the Tweet ID field. We collect the ORT’s and link each ORT with the respective RPT. Thus a chain is formed. We perform this iteratively. In this work we have performed 5 iterations and form a chain of tweets. Each set of Tweet chain is considered as 1 document or file.

We also use additional method for identifying the chain of tweets in the stream of tweets obtained from Twitter API. In this method, we see if the adjacent tweets have same set of twitter handle mentions such as @narendramodi, @BJP4UP etc in the tweets. If there are same set of twitter handle mentions then it is clue for the possibility of the tweets having the same discourse. Such tweets are analysed manually and if found to be having same discourse we make those as chain of tweets. In the data collection we faced an issue that there were many tweets which were just standalone tweets and could not identify the chain of tweets. We have not considered such tweets.

The conversation data was manually created for Malayalam by two native speakers who have are trained in literature and linguistics. The created data is translated to Hindi, English, Tamil by translators and verified by

2.2. Corpus Annotation

The corpus was first Tokenized as it is the initial step in all corpus creation for NLP. Here we used the Tockenizer developed in house for Indian languages. The Tokenized data was given to PALinkA [6], which is an open source tool from University of Wolverhampton. PALinkA is the abbreviation of Perspicuous and Adjustable Links Annotator, an Annotation tool. The corpus was annotated using, PALinkA tool. The texts were annotated using the guidelines which treated all noun phrases (NPs) as markables. It is a language independent tool, written in java. It is tested on Tamil and other Indian Languages, also it is user friendly, we can annotate by selecting the markables and click on it.

The input file to PALinkA has to be a well-formed XML file and the produced output is also a well-formed XML. The pre-processed files with all syntactic information to be annotated should be in XML format. 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. After annotation, these XML files are converted to column format files which are required for the machine learning system.

In this task, the annotators have to mark the referential links between entities in a text. Each anaphor receives a unique ID, and a link between two entities is marked using these IDs. These IDs are automatically managed by PALinkA.

The corpus was annotated by language editors, who were either native speakers or had Masters level education in that language. The corpus for Tamil and Malayalam were annotated by native speakers. English and Hindi corpus was annotated by language editors who had Masters Qualification in those languages.

Each file was annotated by 2 language editors and it was observed that there was good agreement between the annotators. We obtained a kappa score of 0.95 showing good inter annotator agreement.

2.3. The Problems in Annotation

There were certain issues in annotation and they are the split antecedent and No Antecedent. These issues are important issues in anaphora resolution and they are explained below with examples.

i) Split antecedent: Split antecedent is antecedent which consists of more than one NP.

John waited for Maria. They went for pizza.

In the above example they refers to both John and Maria. John and Maria are considered as split antecedent.

ii) Ellipsis: In conversation data there were many elliptical constructions where the antecedent was elided in the discourse. In such cases we annotated the pronoun and the antecedent was manually marked.

2.4. Corpus Statistics

The details of the corpus is discussed in this section. Here we give the number of file, number of tweets, number of anaphors and the anaphors having antecedents. This will give the overall picture of the corpus used for training and testing. The data statistics for all the four languages are given Table 1.

Table 1
Data Distribution of All languages

Description	English			Hindi			Malayalam			Tamil		
	Train	Test	Total	Train	Test	Total	Train	Test	Total	Train	Test	Total
No. of Files	110	100	210	700	85	785	309	61	370	281	147	482
No. of Tweets	875	1000	1855	2096	586	2682	1465	415	1880	1690	1165	2855
Total No. Of Anaphors	765	825	1590	1130	821	1951	845	325	1170	565	426	991
Anaphors having Antecedents	445	535	980	585	425	1010	275	180	455	345	320	665

2.5. Distribution of Anaphors

The distribution of anaphors in each language varies and it is necessary to have a minimum number from each category in the corpus. The below tables give the various types of anaphors and their representation in training and testing corpus.

Table 2
English Anaphors Distribution

Anaphors Type	Train	Test	Total
He	340	396	736
She	108	136	244
It	97	125	222
Others (I,we, they)	220	168	388
Total	765	825	1590

Table 3
Hindi Anaphors Distribution

Anaphors Type	Train	Test	Total
vaha (vo, ve)	226	145	371
Usa (usako, use... etc)	904	676	1580
Total	1130	821	1951

Table 4
Malayalam Anaphors Distribution

Anaphors Type	Train	Test	Total
avan (avanai, avanaaly... etc)	511	168	679
avaly (avallukku.. etc)	289	128	417
Enne	45	29	74
Total	845	325	1170

Table 5
Tamil Anaphors Distribution

Anaphors Type	Train	Test	Total
avan (avanai, avanaaly... etc)	79	54	133
avaly (avallukku.. etc)	67	42	109
avar (avarru)	314	197	511
Atu	105	133	238
Total	565	426	991

3. Task Definition

The task proposed is to develop an anaphora resolution (AR) system from the Twitter data and conversation data annotated for anaphor and its antecedent. The languages considered are from Indo Aryan and Dravidian families and they are Hindi, Tamil and Malayalam respectively. Here we also have given annotated corpus for resource research language English with the view that it could be used for resolving anaphor for less resource language. The corpus was annotated for anaphor and its antecedent. The tokenization was done and the data was in column format.

3.1. Groups Registered

There were 6 groups registered and took the training and test data. Only one group submitted their runs. The details of the groups and their affiliations are given in the Table 6 below.

Table 6
Registered Groups - Enlisted

Sl No	Groups Name	Team members	Affiliation	Language data requested
1.	Vijay Kumari	Hriday Kedia, Vijay Kumari, Yashvardhan Sharma	BITS Pilani	All Languages
2.	Pavan Kandru	NIL	iREL, IIIT Hyderabad	All Languages
3.	Abhinav Kumar	NIL	Shiksha 'O' Anusandhan, Deemed to be University, Bhubaneswar	All Languages
4.	Zengman Kou	NIL	Harbin Engineering University (HEU Harbin, China	English
5.	Yuning Zhang	NIL	Harbin Engineering University (HEU), Harbin, China	English
6.	Bin Wang	NIL	Harbin Engineering University (HEU), Harbin, China	Tamil

3.2. The Group Submitted

The participants were asked to submit their test runs in the format as given in training data. Out of the six groups who took the data, only one group submitted the run for Hindi. Others did not submit the runs.

The group by Ms. Vijay Kumari from BITS Pilani, Pilani alone submitted the run for the language English. They have used a statistical pretrained model of NeuralCoref network by huggingface for English.They have identified the named entities first and then the model is trained for a set of features.This training is done by using a set of intial word embeddings and training them on Ontonotes corpus.

4. Evaluation of Test Run

The evaluation was done using the standard evaluation metrics: Precision, Recall and F-measure.

Table 7
Evaluation Results

Team Name	Language	Precision	Recall	F-measure
Vijay Kumari-BITs Pilani	English	0.30	0.25	0.27

5. Conclusion

We have conducted the task on Anaphora resolution for microblog data from Twitter and conversation data. There were six registrations. The data was given to all the six groups, but only one group submitted the run. Since it is a difficult task many could not submit the runs. In future it is hoped that the data given will help in research in this area.

6. Acknowledgements

Author would like to thank Dr. Pattabhi R K Rao, Mrs. Gracy L and Ms. Padmapriya for developing the corpus.

7. References

[1] Akilandeswari A and Sobha Lalitha Devi. "Anaphora Resolution in Tamil Novels", In Rajendra Prasath, Philip O'Reilly, T. Kathirvalavakumar (Eds), Mining Intelligence and Knowledge Exploration, Springer LNAI Vol 8891 pp. 268-277, (2014).

[2] Akilandeswari A., and Sobha Lalitha Devi. "Conditional Random Fields Based Pronominal Resolution in Tamil", In International Journal on Computer Science and Engineering, vol. 5(6): 601 – 610, (2013).

[3] C Kennedy and B Boguraev. "Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser", in Proc. of the 16th International Conference on Computational Linguistics (COLING'96), Denmark, pp. 113-118. (1996).

[4] R. Mitkov. "Factors in Anaphora Resolution: They are not the only Things That Matter. A Case Study Based on Two Different Approaches", in Proc. of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution, Spain. pp. 14-21. (1997).

[5] S. Lappin and H. Leass. "An Algorithm for Pronominal Anaphora Resolution", Computational Linguistics, vol. 20, 4. pp. 535-561. (1994).

[6] Constantin Orasan. "PALinkA: a highly customizable tool for discourse annotation". In Proceedings of the 4th SIGdial Workshop on Discourse and Dialog, pages 39 - 43, Sapporo, Japan. (2003)

[7] Sobha Lalitha Devi. "Resolving Pronouns for a Resource-Poor Language, Malayalam using Resource-Rich Language, Tamil", In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pp. 611–618., (2019). https://doi/10.26615/978954-452-056-4_072

[8] Vijay Sundar Ram, R. and Sobha Lalitha Devi. "Pronominal Resolution in Tamil ing Tree CRFs", In Proceedings of 6th Language and Technology Conference, Human Language Technologies as a challenge for Computer Science and Linguistics - 2013, Poznan, Poland, LNAI pp. 333-337. (2013).

[9] Sobha Lalitha Devi (2020). 'SocAnaRes-IL20: Anaphora Resolution from Social Media Text in Indian Languages @ FIRE 2020 - An Overview', In the Forum for Information Retrieval and Evaluation-2020, IDRBT, Hyderabad, India.