Automation and Validation of Annotation for Hindi Anaphora Resolution

Pardeep Singh
Computer Science and Engineering
National Institute of Technology
Hamirpur, INDIA

Kamlesh Dutta
Computer Science and Engineering
National Institute of Technology
Hamirpur, INDIA

***Abstract*—The process of labelling any language genre by which one can extract useful information is called annotation. This provides syntactic information about a word or a word phrase. In this paper, an effort has been made to provide the algorithm for semiautomatic annotation for Hindi text to cater anaphora resolution only. The study was conducted on twelve files of Ranchi Express available in EMILLE corpus. The corpus is originally tagged for demonstrative pronouns. The detection of the pronouns is supported by the incorporation of seven tags. However the semantic interpretation of the demonstrative pronoun is not supported in the original corpus. In this paper an effort has been made to automate the process of tagging as well as the handling of semantic information through addition tags. It was conducted on 1485 demonstrative pronouns. The average accuracy of precision, recall and F measure is 74, 71 and 72 respectively.**

Keywords—Annotation; natural language processing; demonstrative pronoun; semantic category; indirect anaphora; semiautomatic annotation

I. Introduction

Natural language processing has attracted the researchers' volition to enhance the natural language resources during the last few decades. A number of applications of natural language processing need the syntactic meanings of words or word phrases. These meanings are used for different applications like, part of speech tagging, information retrieval, text summarizations, question answering, anaphora resolution, etc. So, the annotation will play a pivotal role in these NLP applications. In this study, the systematic discussion has been held of the labelling process of demonstrative pronoun in the context of anaphora resolution.

Anaphora is a process of finding the referring expression in the discourse. Wrong correlation of referring expression in the genre affects all applications of NLP.

Example 1:

"***They*** don't understand why it seems like bad behavior on Wall Street is rewarded, but hard work on Main Street isn't, or why Washington has been unable or unwilling to solve any of our problems", (Obama 2010, http://www.diva-portal.org/smash/get/diva2:531167/fulltext01.pdf, accessed on 18th Aug, 2015).

In example 1, Obama referred to the people in a negative way by implying that they might be inferior, since Obama assumed that they do not understand how the economic crisis was solved. Though the Obama wants to refer the Congress rather people.

Example 2:

"Now, our friends down in Tampa at the Republican Convention were more than happy to talk about everything ***they*** think is wrong with America. But ***they*** didn't have much to say about how ***they'd*** make it right. ***They*** want your vote, but ***they*** don't want you to know their plan. And that's because all ***they*** have to offer is the same prescriptions ***they've*** had for the last 30 years" (Obama, Sept 6, 2012, http://www.presidency.ucsb .edu/ws/index.php?pid=101968, accessed on 18th Aug, 2015.)

In example 2, '***they'*** refer to a specific group, which does not belong to Democrats and this demonstrative pronoun creates serious problems in anaphoric context. Moreover the interpretation of machine has been always error prone.

So, anaphora resolution itself is a significant problem. To address this problem annotation of any corpus is crucial, while formulation, evaluation and optimization of any algorithms in NLP, particularly automation of anaphora resolution. Annotation becomes the prerequisite condition for anaphora and other applications for better accuracy.

II. Background of Annotation

A number of attempts have been made to retrieve the information from the text by a number of means; one of them is annotation. There is no standard annotation scheme which can fulfill all the requirements. The different labeling schemes have been adopted to address the different problems. In this regard the most commonly used practices are phrase structure, dependency, HPSG (Head-driven Phrase Structure Grammar) and Hybrid (Phrase structure and Dependency, both). Penn Treebank [1] is the most used and adapted annotation scheme; firstly for English and then in other languages. A number of languages parsed according to Pen Treebank are like, Arabic, Bulgarian, Chinese, Czech, Danish, Dutch, English, Estonian, Finnish, etc.

III. Motivation

Annotated corpora promise to be valuable for researchers as diverse as the automatic construction of statistical models. Written or spoken language provides the raw data to investigate, evaluation and comparison of different linguistic tools/ models. Combining raw language data with linguistic information, offers a promising basis for the development of new efficient and robust NLP methods. Real world texts annotated with different strata of linguistic information can be used for grammar induction. Annotating the corpora manually takes rigorous effort and competency too. It is better to draw some conclusion/ rules which lead to fully or partially automate this process.

Skill and competency levels of human beings always impose the restriction on the accuracy of annotation. Human interpretation of discourse and its constituents may vary. Understanding may become subjective in the context and may lead to incorrect annotation. Though, manually tagged data are the most preferred practice, may be due to the ultimate benchmark for accuracy of any NLP task, i.e., human interpretation, the need for having automation is necessitated by the availability of voluminous digital data these days and the limitation imposed by the human capacity.

IV. Literature Survey

A number of attempts have been made to label the text or dialogue. Broadly, these are classified under four categories. First dependency structure, second phrase structure, third HPSG (Head-driven Phrase Structure Grammar) and fourth is hybrid labeling. Technically, hybrid type of annotation is to serve two purposes; phrase structure as well as dependency. These annotation schemes are being used to enrich language resources. Multiple corpuses for same language is an added advantage for researchers to test and validate their tool/ technique for the same. It helps to create a more generic solution for those languages.

A. Manually Annotated Corpora

Fully tagged corpus is a necessity of any language to groom its automation tools. A number of corpora are available in English and European languages. The Prague Dependency Tree Bank (PDT) annotated up to three levels; morphological, syntactic annotation and third level were linguistic level [1]. Susanne Corpus as treebanks exist for English [2], the Lancaster Parsed Corpus [3] and another corpus for English, the International Corpus of English [4], the Prague Dependency Treebank for Czech [5]. Treebank projects for other languages made in the recent times, e.g., for French which is tagged for morphosyntax, lemmas, compounds, lexical clusters and phrase boundaries [6], Italian corpora annotated with grammatical relations and syntactic representation of sentence [7], syntactically parsed for Spanish [8], lexically annotated speech corpora of Turkish was an attempt by marking derivation boundaries by [9], and a dependency structured Russian corpora which was lemmatized, morphologically and syntactically tagged [10]. The annotation of the German TIGER Treebank [11] is done in a manner so that it can easily be exported to XML. They consider the verb-sub categorization, coordination, appositions and parentheses as well as proper nouns. TIMEBANK is richly annotated to indicate events, times, and temporal relations [12].

B. Annotated corpora for Coreference and Anaphora

There are a few corpora which were annotated, especially for anaphora resolution or co-reference resolution. Lancaster Anaphoric Treebank [13] of Associated Press with 100000 words and annotated according to UCREL annotation scheme. This was the joint venture of UCREL and IBM. MUC-6 and MUC-7 annotate co-referential link of 65000 words. Similar to MUC scheme a tool ClinKa was developed to annotate English genre at University of Wolverhampton [14]. Another corpus developed by the members of University of Stendahls Grenoble and Xerox Research Center Europe [15] by creating an anaphoric and cataphoric link. It addresses the zero noun anaphora, adverbial anaphora, indefinite pronoun, demonstrative pronoun, third person personal pronoun, and personal pronoun. French corpus [16], annotate anaphoric links in MUC. Few multilingual corpora are available like English-Romanian corpus [17], technical manual of English- French for co-referential link at the University of Wolverhampton.

C. Automatic or Semiautomatic Annotated Corpora

An attempt for corpus annotation for labelling semantic and syntactic meaning of word and word phrases for coverage of deep parser to generate syntactic structure, semantic representation and discourse information for dialogue by means of semiautomatic technique [18]. The EPAC was a speech corpus, it consists of a set of 100 hours of conversational speech manually transcribed [19]. This spoken corpus automatically annotated by automatic segmentation, transcription, POS tagging and other tools. The Diachronic German Corpus [20] was automatically annotated by a suite of NLP tools. These tools are integrated into WebLicht and CLARIN-D. WebLicht Service Oriented Architecture is used as an integrated environment. A corpus is trained with automatic system for semantic [21]. It advocates coreference, quantification, and defined a set of semantic rules for many other higher-order phenomena, which was left out by Penn Treebank.

D. Annotation for Hindi

Botley & Mc Energy [22] proposed an annotation scheme for English to resolve anaphora. Later this scheme modified by Botley [23] again for the same purpose and same language with emphasis on indirect anaphora. They considered the recoverability of antecedent, direction of reference, phoric type, syntactic function, antecedent type to annotate three genres. These corpora are the American Printing House for the Blind (APHB) Corpus, the Associated Press (AP) Corpus, and the Hansard Corpus [23]. Recoverability refers to the relation between referring expression and its corresponding antecedent in context of demonstrative pronoun and this feature based on Halliday and Hassan [24]. Feature “Phoric type” is distinction between substitution and reference [24]. Their tag set adapted to annotate Hindi by Sinha [25], and later [26] added a few more tags. Reference [27] used some semantic information for indirect anaphora categorization. The author has identified a few semantic categories to classify indirect anaphora in Hindi.

V. Methodology

A. Tag set used

The demonstrative pronouns are understood in terms of an unordered paradigmatic set of five distinctive features [22] . Syntactic Function and Antecedent Type, two other features, which are proposed by [23]. Last three features in table 1 facilitate to identify indirect anaphora [27], [26], [25]. Recoverability of antecedent helps to identify the same.

TABLE I. FEATURE USED FOR ANNOTATION

No. of Feature	Feature	Value1	Value2	Value3	Value4	Value5
1.	Distance Marking	P (proximal)	D (Distal)	None	None	None
2.	Nature Of deixis	P (Pronoun)	D (Demonstrative)	Z (Zero)	None	None
3.	Recoverability of Antecedent	D (Directly Recoverable)	I (Indirectly Recoverable)	N (Non-recoverable)	0 (not applicable, e.g. exophora)	None
4.	Direction of reference	A (anaphoric)	C (cataphoric)	0(not applicable, Exophoric or deictic)	None	None
5.	Phoric Type	R (Referential)	0 Not Applicable	None	None	None
6.	Syntactic Function	M (Noun Modifier)	H (Noun Head)	0 (Not Applicable)	None	None
7.	Antecedent Type	N (nominal)	P (propositional/ Factual)	C (Clausal)	J (Adjectival)	O (None)
8.	Pronoun pattern	Pronoun and subsequent construct in the sentence				
9.	Case marker/ Connective	Case marking or connective following the pronoun				
10.	Semantic/ category	Semantic categories				

B. Corpus selection

This study has been conducted on EMILEE corpus. The Hindi written corpus contains a total of approximately 12,390,000 words in Unicode. It is pre annotated corpora for demonstrative pronouns. One component of this corpus is based upon Ranchi Express news items. In pre annotated corpus there are seven features. We have modified this corpus with an annotation scheme according to [27]. Being in Unicode is an added advantage of using EMILEE corpus.

Example 1 is manually annotated Ranchi Express news from EMILEE corpus according to Botley's annotation. In this annotation only seven tags are there and their respective value of उन्होंने (unhone) pronoun is DPDARHN.

Example 3:

<body>

<p>किसी मंत्री को

हटाने का सवाल नहीं : मरांडी</p>

<p>रांची: मुख्यमंत्री बाबूलाल मरांडी ने आज
विधानसभा में कहा कि पलामू में एक लड़की के अपहरण की घटना के क्रम में झारखंड मंत्रिमंडल से किसी सदस्य को हटाने का सवाल ही पैदा नहीं होता। <w tag = " DPDARHN"> उन्होंने</w> कहा कि <w tag= " P D DARMN">यह</w> मामला कई दिनों से चर्चा में है लेकिन, घटना अपहरण की है अथवा लड़का और लड़की स्वेच्छा से गए हैं <w tag="PDDARHC">यह</w> जांच का विषय है।

We have considered the data from EMILLE corpus. We picked one segment of corpus which is based on the news items from Ranchi express. In this study, we analysed twelve files of plain text.

Example 4: Manually annotated Ranchi Express news from EMILEE corpus according to table 1.

<body>

<p>किसी मंत्री को

हटाने का सवाल नहीं : मरांडी</p>

<p>रांची : मुख्यमंत्री बाबूलाल मरांडी ने आज विधानसभा में कहा कि पलामू में एक लड़की के अपहरण की घटना के क्रम में झारखंड मंत्रिमंडल से किसी सदस्य को हटाने का सवाल ही पैदा नहीं होता। <w tag= "D, P, D, A,R,H,N,उन्होंने,यह,_,_,_">उन्होंने</w> कहा कि <w tag="P,D,D,A,R,M,N,यह, यह,_,_,_"> यह</w> मामला कई दिनों से चर्चा में है लेकिन, घटना अपहरण की है अथवा लड़का और लड़की स्वेच्छा से गए हैं <w tag="P,D,D,A,R,H,C,यह,_,_,_">यह</w> जांच का विषय है।

In example 4 additional three tags have been attached उन्होंने,यह, ,_,_,_". First is pronoun pattern, second case marker and third semantic category. In this example null is denoted by '_'.

C. Algorithm

Though, some researchers advocated syntactic features of languages for annotation. Reference [27] suggested some specific pattern of pronoun and other words which is categorized in pattern. Secondly, case marker elaborates the pronoun significance and binding with its antecedent. This algorithm annotate only last three tags discussed in example 4.

```
Step1: Input the set of case marker.
Step2: Input pattern for pronouns
Step3: WHILE(file in the lists of files) REPEAT
        I)   Find "<w tag, and corresponding ">"
           a.  Extract the feature list call it tag
        II)  Split the tag list into list of individual features
        III) Generate_Case_Marker_and_Pronoun_Pattern()
           a.  Define window size for pattern
```

```
            b. Within window size, find the
               case_Marker and pattern_follow_Pronoun
            c. Extract case_Marker and Pattern_follow_
               Pronoun
        IV) Classification()
            a. Extract semantic category, which is tag
               [10], Pronoun, which is tag [8], pattern
               following pronoun which is stored in
               string tag [].
            b. Sequential search these in the rule based
               classifier.
            c. Output the CLASS; return CLASS.
    END WHILE
Step4: Print all the text along with modified tag to files.

Step5: Stop
```

EMILEE corpus is annotated with seven tags. Last three tags (Pronoun pattern, Case marker/ Connective, Semantic category) have to be annotated either manually or automatically. Reference [27] draw some rules for classification of indirect anaphora in demonstrative pronoun.

As a first step, the eight case markers of Hindi given as input then pronoun pattern. Third step is to read all files to be annotated. It extracts all tags (feature list) which is already annotated in EMILLE corpus. Then it generates the case marker and a pronoun pattern by defining window size. We took window size five. In this window, algorithm will search specific pronoun pattern and case marker according to [27]. Pronoun was matched with the given list of pronoun and proceeding word; which indicates recoverability of antecedent on the basis of semantics. We considered the first hit for pronoun pattern and case marker. Then extract the case marker and pronoun pattern. In fourth step syntactic category, pronoun and pattern following pronoun are stored as the element of a string tag (i.e. tag [8], tag [9] and tag [10] respectively). Then apply the rules given by [27]. In '*b*' part of '*fourth'* step of above algorithm (i.e. Sequential search these in the rule based classifier) has been adopted from [27]. The output is stored in one class. This is the required output with file for all ten tags.

D. Case marker used

Hindi consists of nine case markers. First eight are in use and the last one "hey" (हे) is obsolete, or less in use practically. These case markers specify the binding with anaphor and antecedent. In ergative case marker there are few bindings and exceptions, e.g. the postposition "ne" (ने) must come right after the subject; the subject changes in oblique the perfect form of the verb now agrees with the direct object in number and gender. In case of above condition, number and gender agreement puts the bindings between verb and direct object. Exceptions are:

- If the object is not stated, or if the object is followed by को (ko) then the perfect form of the verb should be in masculine singular form.
- The auxiliary verb (if any) also agrees with the object, not the subject.

The eight case markers are ergative, nominative, ablative, accusative, instrumental, genitive, dative and locative. We have considered these eight cases of Hindi for the study. Few cases are given below, though in linguistic perspective, these cases come as a suffix.

- ne – it marked as ergative marks the subject or topic (but only in the past perfective tense for transitive verbs)
- ka/ke/ki - marks the genitive
- ko - marks the accusative or dative, - typically means "from" or "by", also marks the passive agent
- mein, par - locative "in", "at"

E. Accuracy measurement of automated tags

In natural language processing mainly two metrics are used. The First precision and the second recall. Another metrics is derived from precision and recall, which is called F measure

- Precision (P) is the fraction of retrieved documents that are relevant

$$p = \frac{\#(relevant\ item\ retrieved)}{\#(retrieved\ items)} \tag{1}$$

- Recall (R) is the fraction of relevant documents that are retrieved

$$R = \frac{\#(relevant\ item\ retrieved)}{\#(relevant\ item\ )} \tag{2}$$

These notions can be made clearer by examining the following contingency table:

TABLE II. CONTIGENCY TABLE

	Relevant	Non relevant
Retrieved	True positive(tp)	False positive(fp)
Not retrieved	False negative(fn)	True negative(tn)

$$P = \frac{tp}{(tp+tf)} \tag{3}$$

$$R = \frac{tp}{(tp+fn)} \tag{4}$$

There is another alternative to calculate accuracy of data set. It is the ration of true selection of text and the sum of all selections (all true + all false).

$$Accuracy = \frac{tp+tn}{tp+fp+fn+tn} \tag{5}$$

This seems plausible, since there are two actual classes, true and false, and an information retrieval system can be considered as a two-class classifier which attempts to label them as such. We are using only equation first, second and seventh.

- F measure : It is the harmonic mean of Precision and Recall

Given n points, x1, x2,…………x_n, the harmonic mean is:

$$\frac{1}{H} = \frac{1}{n}\sum_{i=1}^{n} 1/x_i \tag{6}$$

So, the harmonic means of precision and recall is:

$$\frac{1}{F} = \frac{1}{2}\left(\frac{1}{R} + \frac{1}{P}\right) = \frac{P+R}{2PR} \tag{7}$$

With the help of above equation (7), we will calculate the F measure. The authenticity of results checked against only three metrics, i.e. precision, recall and F measure.

VI. Result and Discussion

The study was conducted on twelve files of Ranchi Express from EMILEE corpus. It consists 206 news items, and 1485 demonstrative pronouns. An effort has been made to automate tagging of last three tags given in table 1. A set of rules are applied to accomplish the task. Three accuracy metrics have been considered; precision, recall and F measure. The Table 3 shows all the twelve files and number of pronouns found in the respective file. Number of news and length of news per file is directly proportional to the number of pronouns.

TABLE III. Pronoun Count Per File

File No.	Number of Pronoun / file
1.	61
2.	148
3.	101
4.	108
5.	104
6.	155
7.	138
8.	128
9.	102
10.	111
11.	151
12.	178
Total	1485

In each file for each pronoun performance metrics are calculated. In the first file, first pronoun was “iss” (इस) and the value of precision, recall & F measure is 66.66, 66.66 and 66.66 respectively. These values for second pronoun are 100, 100 and 100. The average precision, recall and F measure of file 1 at serial number one in the table. Hence, the average of each metric is calculated and given in the table 4 below along with its respective file. Then again average of all files is calculated. Precision varies from 65.52 to 79.08 and recall 65 to 75. This depicts there is no major variation in all metrics.

In figure 1, the annotated genre; on the axis X number of files are given. On axis Y, percentage of accuracy of three tags in the terms of precision, recall and F measure is given. Though the pattern of result of all these three parameters for all the twelve files is almost the same. The average accuracy of precision, recall and F measure is 74, 71 and 72 (approximately) respectively.

A. Error Analysis

We classify the entire results into three types; worst, average and best results across the data set (all files) and their pronoun.

TABLE IV. Precesion, Recall and F Measure Per File

S. No.	File Name	Precision	Recall	F measure
1.	File1	77.22	75.66	76.3
2.	File2	78.83	75.30	76.79
3.	File3	66.81	71.61	68.82
4.	File4	76.94	73.44	74.84
5.	File5	78.96	74.98	76.67
6.	File6	74.24	70.22	71.93
7.	File8	68.51	65.02	66.44
8.	File9	72.70	68.34	70.11
9.	File10	76.89	72.24	74.17
10.	File11	79.08	75.53	76.97
11.	File12	73.70	69.20	70.96
12.	File13	65.52	61.94	63.30
Average of each metrics		**74.12**	**71.12**	**72.28**

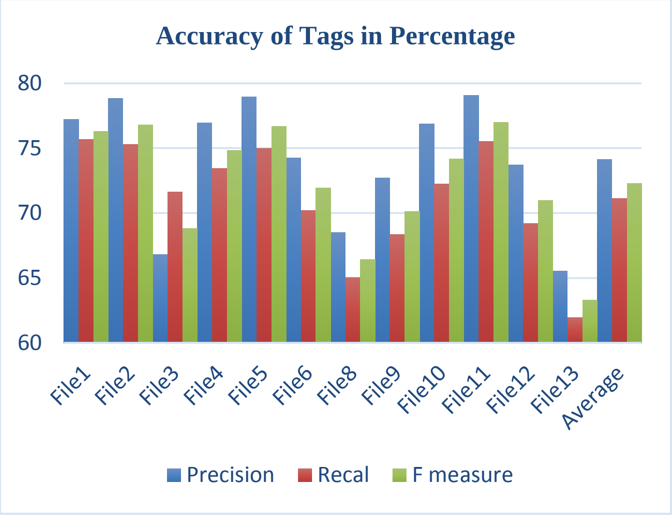

Fig. 1. Accuracy of tags

*1) **Worst case (file 8, pronoun 81st):*** It is observed that in the 7th file (file name is file 8) having 81st to 83rd pronoun produced the result zero for precision, recall and F measure**.** Automated file has been given below. Special cases, which annotate genre with zero accuracy compared with manual tagged file for the same text with same pronoun. In the proposed algorithm, first, we stored all case marker/ connector and pronoun patterns. The algorithm search for the required case marker and pronoun pattern. Wherein window size is more than one in which we are seeking the case marker and the pronoun patterns.The precision and recall was calculated of the entire string after the 7th tag. In manual tagging there are four entries, including pronoun after the 7th tag. On the other hand automatically tagged 81st pronoun has ten entries after the 7th tag. Last three tags are same in automatic annotation and manually tagged annotation for this example.

Pronoun 81st : *अगर वास्तव में <w tag="D,D,D,A,R,M,N,उन,इन,उन्होंने,उनका,वे,उसी,उन,_,_,_">उन</w> लोगों ने छेड़छाड़ की घटना के मुद्दे पर ही*

Pronoun 82nd : <w tag="P,D,D,A,R,M,N,इन,उन्होंने,उनका,वे,उसी,उन,_,_,_">इन</w> सिख युवकों की हत्या की थी तो निश्चित रूप से

Pronoun 83rd : <w tag="D,P,D,A,R,H,N,उन्होंने,उनका,वे,उसी,उन,_,_,_">उन्होंने</w> योजना बनाकर <w

The pronoun 82nd and 83rd of the same file depict the same result due to over length entries in string of case marker and pronoun pattern. It can be fixed with the help of window size. File 3, file 8, file 12 falls in the category of the worst case. Though the average of these files is 61 to 75 percentage.

Manually tagged pronoun for comparison of file 8.

Pronoun 81st : अगर वास्तव में <w tag="D,D,D,A,R,M,N,un,null,null,null">उन</w> लोगों ने छेड़छाड़ की घटना के मुद्दे पर ही

Pronoun 82nd <w tag="P,D,D,A,R,M,N,inn,null,null,null">इन</w> सिख युवकों की हत्या की थी तो निश्चित रूप से

Pronoun 83rd : <w tag="D,P,D,A,R,H,N,unhon-ne,null,null,null">उन्होंने</w> योजना बनाकर

2) **Average case (file 1, pronoun 1st):** It was considered as average case if the accuracy of precision, recall and F measure is fifty percent or more. File 1, file 4, file 6, file 8, file 9 and file 11 lies in the average case.

Manually annotated file1

<body>

<p>बालूमाथ, १६ मार्च: आज शाम बालूमाथ थाना अंतर्गत लेबडाही जंगल में प्रतिबंधित एम.सी.सी. संगठन के उग्रवादियों और पुलिस के बीच मुठभेड़ हुई।

Pronoun 1st : <w tag="P,D,D,A,R,M,N,iss,null,null,act">इस</w> मुठभेड़

Automatic tagged file1

<body>

<p>बालूमाथ, १६ मार्च: आज शाम बालूमाथ थाना अंतर्गत लेबडाही जंगल में प्रतिबंधित एम.सी.सी. संगठन के उग्रवादियों और पुलिस के बीच मुठभेड़ हुई।

Pronoun 1st : <w tag="P,D,D,A,R,M,N,इस,_,_,_">इस</w> मुठभेड़

In the above example of file 1 and pronoun 1st, only one is/has mismatched tag. Tag 10 differs in manual and automatic tagging.

3) **Best Case (file 1, pronoun 2nd):** It was considered as the best case if all three metrics have 100 percent accuracy. In this example null is equal to _ (underscore) sign. Each field matched and in both files.

Manually annotated file1

Pronoun 2nd : <w tag="P,D,D,A,R,M,C,iss,null,null,null">इस</w> गोलीबारी में पुलिस दस उग्रवादियों को मार गिराने का दावा कर रही है। उग्रवादियों की गोली से बालूमाथ थाना पुलिस वाहन के चालक अशोक कुमार (३५) गंभीर रूप से घायल हो गया है। पुलिस और एम.सी.सी. के बीच ढाई घंटे तक मुठभेड़ हुई है।

Automatic tagged file1

Pronoun 2nd : <w tag= "P,D,D,A,R, M, C, इस, _,_,_">इस</w> गोलीबारी में पुलिस दस उग्रवादियों को मार गिराने का दावा कर रही है। उग्रवादियों की गोली से बालूमाथ थाना पुलिस वाहन के चालक अशोक कुमार (३५) गंभीर रूप से घायल हो गया है। पुलिस और एम.सी.सी. के बीच ढाई घंटे तक मुठभेड़ हुई है।

B. Inference:

- It reflects that the seeking window should be decreased to the optimal size in order to avoid additional case marker and pronoun pattern. This particular example (**Worst case, file 8, pronoun 81st**) depicts that the scenario of case marker string and pronoun pattern string may have multiple entries due to the number of pronoun and the number of case marker come consecutively in discourse. e.g. उन (un), इन(in), उन्होंने(unhone).
- The last three features was automated. There are two methods to match ten features. First, start the matching of features from 7th tag to 10th tag. And the second is to match the last three features of automatic and manual files. Then discard the additional entries between 7th and last but 3rd. It will solve the problem of additional entries in case marker and pronoun string. It also will improve the accuracy in the terms of precision, recall and F measure.
- Refining the number of rules will increase the accuracy of automatic annotation. These rules define the pattern of case marker/ connectives and pronoun. These two features have more contribution for error.

VII. Conclusion and Future Work

A. Depiction of results

Few results were concluded from ongoing work. Before arriving at the conclusion, twelve files are studied of monologue, and 206 news items which consist 1485 pronouns. Accuracy varies from 65 to 79 for precision. Recall varies 62 to 75 and F measure has been maximum and the minimum values are 77 and 63 percent respectively. Average of precision, recall and F measure remained 74, 71 and 72 percent. File number twelve (12) has the lowest accuracy for all the three metrics and file ten (10) has the highest.

B. Conclusion

This dataset depicts the generalized picture of the genre. Fine tuning of rules besides considering few other semantic categories increase the accuracy of results. In the proposed algorithm the window size is five or till the end of sentence. It will count all the pronoun patterns and add to the tag set. It will make additional entries in values of feature set string. The remedial action for window size is to decrease it to one. Now, there are few advantages and few disadvantages of window size one.

Decreasing window size may increase the accuracy in terms of precision, recall and F measure. While the disadvantage is that it will skip further potential pronoun pattern which may yield higher accuracy. That means we have to go for the best hit. To find the best hit one has to develop new logic. Potential pronoun pattern and case marker may be tested for other genres. It also optimizes the results. In this work we are replacing the potential pronoun pattern with the first hit. In future it may be replaced with the best hit by other logic. Fine tuning the rules and taking gold standard dataset increase the percentage of accuracy.

C. Future work

Though the corpus is small. A larger and gold standard dataset may produce more authentic results. Selection and fine tuning of rules increase the accuracy of results. Developing the logic for the best hit in seeking window for potential pronoun pattern and case markers also helps to improve the results. These few issues can be addressed in future work which optimize the results.

ACKNOWLEDGMENT

The author would like to acknowledge Pratheek Adidela for discussion on python libraries, their use and detailed discussion on the logic of indirect anaphora and the sequence of rules to be applied in the given algorithm.

REFERENCES

[1] M. Marcus, B. Santorini and M. A. Marcinkiewicz, "Building a Large Annotated Corpus of English: The Penn Treebank," Association for Computational Linguistics, pp. 313-330, 1993.

[2] G. Sampson, English for the computer, The SUSANNE corpus and analytic scheme, Oxford, UK: Clarendon Press, 1995.

[3] G. Leech, "The Lancaster Parsed Corpus," ICAME Journal, vol. 16, no. 124, 1992.

[4] S. Greenbaum, Comparing English worldwide: The International Corpus of English, Oxford, UK: Clarendon Press, 1996.

[5] J. Hajic, "Building a Syntactically Annotated Corpus: The Prague Dependency Treebank," in Issues of valency and meaning, Karolinum, Praha, 1998, pp. 106-132.

[6] A. Abeill´e, L. Cl´ement and A. Kinyon, "Building a treebank for french," in Proceedings of the Second International Conference on Language , Athens, Greece., 2000.

[7] C. Bosco, V. Lombardo, D. Vassallo and L. Lesmo, "Building a treebank for italian: A data-driven annotation schema," in Proceedings of the Second International Conference on Language Resources and Evaluation LREC-2000, Athens, Greece, 2000.

[8] A. Moreno, R. Grishman, S. L´opez, F. S´anchez and S. Sekine, "A treebank of spanish and its application to parsing," in Proceedings of the Second International Conference on Language Resources and Evaluation LREC-2000, Athens, Greece, 2000.

[9] K. Oflazer, D. Hakkani-T¨ur and G. T¨ur, "Design for a turkish treebank," in proceedings of the Workshop on Linguistically Interpreted Corpora LINC-99, Bergen, Norway, 1999.

[10] I. Boguslavsky, S. Grigorieva, N. Grigoriev, L. Kreidlin and N. Frid, "Dependency treebank for russian: Concept, tools, types of information," in 18th International Conference on Computational Linguistics COLING-2000, Saarbr¨ucken, Germany, 2000.

[11] S. Brants and S. Hansen, "Developments in the TIGER Annotation Scheme and their Realization in the Corpus," in proceedings of the Third Conference on Language Resources and Evaluation LREC-02, 2002.

[12] P. James, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer and D. Radev, "The timebank corpus," Corpus linguistics, pp. 40-48, 2003.

[13] G. Leech and R. Garside, "Running a Grammar Factory: The Production of Syntactically Analysed Corpora or Treebanks," English Computer Corpora: Selected Papers and Research Guide, pp. 15-32, 1991.

[14] R. Mitkov, R. Evans, C. Orasan, C. Barbu, L. Jones and V. Sotirova, "Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies," in Proceedings of DAARC 2000, UK, 2000.

[15] A. Tutin, F. Trouilleux, C. Clouzot, E. Gaussier, A. Zaenen, S. Rayot and G. Antoniadis, "Annotating a large corpus with anaphoric links," in Third International Conference on Discourse Anaphora and Anaphor Resolution DAARC2000, United Kingdom, 2000.

[16] A. P. Belis, L. Rigoust, S. Salmon-Alt and L. R, "Online Evaluation of Coreference Resolution," in LREC 2004 Fourth International Conference on Language Resources and Evaluation, 2004.

[17] S. M. Harabagiu and S. J. Maiorano, "Multilingual coreference resolution," in Proceedings of the sixth conference on Applied natural language processing, Morristown, NJ, USA, 2000.

[18] M. D. Swift, M. O. Dzikovska, J. R. Tetreault and J. F. Allen, "Semi-automatic syntactic and semanticcorpus annotation with a deep parser," in Fourth International Conference on Language Resources and Evaluation LREC-2004, 2004.

[19] Y. Esteve, T. Bazillon, J. Y. Antoine, F. Bechet and J. Farinas, "The EPAC corpus: Manual and automatic annotations of conversational speech in french broadcast news," in proceedings of the seventh conference on International Language Resources and Evaluation(ELRA), Valletta, Malta, 1686-1689.

[20] E. Hinrichs and T. Zastrow, "Automatic Annotation and Manual Evaluation of the Diachronic German Corpus TüBa-D/DC," in Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), 2012.

[21] M. Palmer, D. Gildea and P. Kingsbury, "The Proposition Bank: An Annotated Corpus of Semantic Role," Computational Linguistics archive, vol. 31, no. 1, pp. 71-106, 2005.

[22] S. P. Botley and A. McEnery, "Demonstratives in English: a corpus-based study," Journal of English Linguistics, vol. 29, pp. 7-33, 2001.

[23] S. P. Botley, "Indirect anaphora: Testing the limits of corpus-based linguistics," International Journal of Corpus Linguistics, vol. 11, pp. 73-112, 2006.

[24] M. Halliday and R. Hasan, Cohesion in English, London: Longman, 1976.

[25] S. Sinha, "A Corpus-based Account of Anaphor Resolution in Hindi," UK, 2002.

[26] R. Prasaad, E. Miltaski, A. Joshi and B. Webber, "Annotation and Data Mining of the Penn Discourse TreeBank," in ACL Workshop on Discourse Annotation, 2004.

[27] K. Dutta, S. Kaushik and N. Prakash, "Machine Learning Approach for the Classification of Demonstrative Pronouns for Indirect Anaphora in Hindi News Items," The Prague Bulletin of Mathematical Linguistics, pp. 33-50, 2011.