Sensitivity Analysis of Feature Set Employed for 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
Sensitivity analysis is the process of doing a systematic review involving a sequence of parameter, feature set and decisions to calculate the impact of these parameters on the study. It will guide the researchers to evaluate the parameter to consider their relevance in the study. In this paper we consider two features out of seven tags which were employed to resolve the anaphora in Hindi. These tags and their values analyzed empirically for the corpus. We analyzed 165 news items of Ranchi Express from EMILEE corpus of plain text. It consists 1745 sentences. Eight files of dialogue base from the same corpus have been analyzed which will have 1521 sentences. We exploited tag set proposed by different authors and their features.

General Terms
Natural language Processing, Machine Translation, Issues in machine translation.

Keywords
Coreference resolution; sensitivity analysis; Anaphora resolution; Annotation.

1. INTRODUCTION
Sensitivity analysis is the key to quantitative assumptions, estimates and decisions which are changed systematically to assess their effect on the final outcome. This analysis evaluates and quantifies the impact of each feature on system/framework to calculate the critical factors, overall risk and identification of weightage of feature. It offers the contingency analysis and uses qualitative assumptions for different scenarios.

Machine translation has been a challenging task. This task attracted the attention of researchers after a few decades since the inception of the computer. It involves a number of issues like semantic analysis, syntactic analysis, morphology, word order of language, word sense disambiguation, discourse knowledge, anaphora resolution, etc. All these issues are required to be addressed to increase the accuracy of machine translation. Therefore, resolving anaphora is equally important for translation. It is required that the sensitivity analysis should be carried out for all features used to resolve anaphora.

2. ANAPHORA RESOLUTION
Anaphora resolution is a device to find the referent and referring expression in the sentence or across the sentences.

It can be divided into two parts:-

2.1 Intra-sentential
Anaphora and its antecedent when within the sentence, is called intra-sentential resolution.

Example 1:
Radha ate bananas because **she** was hungry.

In example 1 '**she**' refer to '**Radha**' and both referent and referring expression are in the same sentence.

2.2 Inter-sentential
When referring expression and referent scattered across the discourse, then it is called inter sentential anaphora.

Example 2:

Sachin Tendulkar $_{*(i, j, k, l, m, n)}$, **the master blaster**$_{i}$ has been the most complete batsman of **his**$_{j}$ time. **His**$_{k}$ batting is based on the purest principles: perfect balance, economy of movement, precision in stroke-making, and that intangible quality given only to geniuses: anticipation. If the **biggest cricket icon**$_{l}$ doesn't have a signature stroke - the upright, back-foot punch comes close - it is because **the most prolific run maker**$_{m}$ is equally proficient at each of the full range of orthodox shots (and plenty of improvised ones as well) and can pull them out at will. There are no apparent weaknesses in **his**$_{n}$ game. (http://www.espncricinfo.com/india/content/player/35320.html)

In the above example; words 'Sachin Tendulkar', 'the master blaster', 'his', 'biggest cricket icon', 'the most prolific run maker' refers to only one person Sachin Tendulkar, therefore all these words are co-referential.

There are many instances in the genre when Antecedent is explicit. This is called direct anaphoric expression. And when antecedent is implicit, it is called indirect anaphora. It is a two-fold classification of anaphora. It gives birth to another type of classification of Anaphora resolution.

2.3 Direct Anaphora
In English, anaphora is the reference to the preceding part of the utterance and can be realized by many different linguistic markers, such as pronouns or demonstratives, as we can see from the following examples:

Example 3:

Angel didn't attend college because **she** felt sick.

Example 4:

The monkey took **the banana** and ate **it**.

In the examples (3) and (4) above, the person or entity being referred to by the pronoun, the antecedent, is easily recoverable from the preceding context, therefore, these examples are called direct anaphora, wherein the anaphor and antecedent are co-referential. Here, a reader or hearer would have no trouble to identifying the antecedent, as the nature of

the link between the anaphor and antecedent is fairly straightforward.

2.4 Indirect Anaphora

Indirect anaphora can be thought of as co-reference between a word and an entity implicitly introduced in the text before as we can see from the examples (5) and (6):

Example 5:

"In 1973 the government met the premiers of the western provinces. Just the other day we received copies of an update from the Prime Minister's address to Premier Barrett on the event of the recent conference of western premiers. Some of **that** process is worthy of commendation, which I sincerely extend to the Prime Minister."

Example 6:

Mary was fired.

a) That happened last week

b) That is true

c) That surprised me

In both (5) and (6), the antecedent of 'that' is more difficult to define directly because the antecedent in these cases is not a surface noun or noun phrase, and the link between them is not one of co-reference. Also, the nature of the anaphoric link in these cases means that a reader or hearer may have to carry out a somewhat complex process of inference to arrive at the antecedent. Therefore, these examples can be said to fall under *indirect anaphora* or IA.

3. FEATURE SET SELECTED

Coreference occurs when multiple expressions in a sentence or document refer to the same thing; or in linguistic jargon, they have the same referent. For example, in the sentence; *Radha said she would help me*, 'she' and 'Radha' are most likely referring to the same person or group, and in that case they are co-referent. Similarly, in *I saw Raj yesterday. He was fishing by the lake*. 'Raj' and 'he' are most likely co-referent.Additional information inserted in the text to process any corpus is called tags. A set of tags chosen to process the text for a particular task is called annotation scheme. While, the number of tags used for that particular task are called feature set.

A number of annotation schemes are available for different tasks. These tag set is defined by different authors [3], [5-7] in English, Europium languages and modified for other languages like Turkish, German, Dravidian languages etc.; to create an annotated corpus. There are six features proposed to annotate demonstrative pronoun for English language [6]. The author considers the recoverability of antecedent, direction of reference, phoric type, syntactic function, antecedent type to annotate three genre. These corpora are the American Printing House for the Blind (APHB) Corpus, the Associated Press (AP) Corpus, and the Hansard Corpus [4]. Later, three tags were suggested and adapted the annotation scheme for Hindi [6], [8]. A machine learning approach is proposed for classification of indirect anaphora and added one more tag to previous work [5]. This tag considers the semantic category. The author proposed that apart from some syntactic constraints semantic collocation pattern is also significant feature for indirect anaphora in Hindi [2]. An annotated corpus by adopting the lexically grounded approach of the Penn Discourse Treebank (PDTB) [6], they present a preliminary analysis of discourse connectives in a small corpus scheme. A number of attempts have been made for manual annotation and semiautomatic/ automatic annotation [11-15]. Word order imposes more constraints [1], [9]. The five features[1] of the annotation are systematically eliminated from the study. These features are type of Recoverability of Antecedent, Direction of Reference, Phoric type, Syntactic Function and Antecedent Type. Another study was carried out for the analysis of the future of anaphora resolution [10]

Table 1. Feature Set used for annotation

Feature	Value1	Value2	Value3	Value4
Distance marking	P(proximal)	D(distal)	None	None
Nature of deixis	P (Pronoun)	D (Demonstrative)	Z (Zero)	None

Example 6:

एक सवाल के जवाब में सी.बी.आइ. के अपर निदेशक श्री विश्वास ने स्पष्ट किया कि आभियुक्तों के विरुद्ध गैरजमानतीय वारंट जारी होने के बाद सी.बी.आइ. अगली कार्रवाई शीघ्र करेगी। <s tag="ne,s"><w tag= " D P D A R H N" >उन्होंने</w></s> लालू प्रसाद समेत अन्य आभियुक्तों की गिरफ्तारी की संभावना से इंकार नहीं किया और <w tag="PDDCRHC">यह</w> भी कहा कि <w tag="PDDARMN">इस</w> मामले में गिरफ्तार <o tag="ko,o">आभियुक्तों</o> कोरांची में ही रखा जाएगा। </p>

Example 6 is a case of specific annotation with seven tags proposed by [3-4]. This annotation is already done in EMILEE corpus. First the features was extracted from annotated corpus. Frequency of values of each feature has been obtained as shown in table1 and parentage was calculated to draw general picture features of the corpus.

3.1 Feature set selection

We are using EMILLE corpus. In this corpus each occurrence of demonstrative pronoun is coded in such a manner so that it could be extracted. The pronoun marked as a direct or indirect, does not specify what actually distinguishes direct anaphor from the indirect ones. The corpus is annotated for anaphora using scheme based on [4] and customized for Hindi corpus by reference [5]. In this study, we are considering only four features.

1) Distance marking
2) Nature of deixis

3.1.1 Distance marking

This feature has two values P (proximal) and D (Distal Remaining three values are irrelevant and represented as zero (0). It describes the feature of antecedent that is proximal or distal. In this exercise we calculated the frequency of these values for this feature and calculated the percentage of occurrence.

3.1.2 Nature of deixis

The nature of deixis has three values P (pronominal), D (Demonstrative) and Zero (0). It reveal whether the anaphor is pronominal, demonstrative or zero.

4. RESULT AND DECISION

Analysis has been carried out on 165 news items of Ranchi Express from EMILEE corpus for both monologue and dialogue. This corpus is available in the public domain which provides free license for academic studies. Seven tags are

already there in the corpus. The corpus is tagged according to table no 1 annotation scheme with additional tag of case marker and subject /object.

<body>

किसी मंत्री को

हटाने का सवाल नहीं : मरांडी

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,unhon-ne,null,null,null"> उन्होंने </w> कहा कि <w tag="P, D, D, A, R, M, N, yeh, maamla, null, event" >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, yeh, null, null, null" >यह</w> जांच का विषय है। मुख्यमंत्री प्रश्नकाल के दौरान कांग्रेस के विधायक चंद्रेश्वर दूबे ने मामला उठाते हुए कहा कि राजस्व मंत्री के पुत्र द्वारा एक लड़की के अपहरण की घटना के आलोक में मुख्यमंत्री क्या राजस्व मंत्री को कैबिनेट से हटाना चाहते हैं। <w tag="D, P, D, A, R, M, N, unkaa, null, null, null">उनका</w> कहना था कि १० मार्च को मुख्यमंत्री श्री मरांडी राजस्व मंत्री के आवास पर गये थे <w tag="D,D,D,A,R,H,N,uss-ke,null,ke,null">उसके</w> बाद १३ मार्च को अपहरण की घटना हुई।

4.1 Distance Marking

Distance marking tag elaborates the antecedent. It has two values which indicate whether antecedent is distal or proximal. Third one is zero value for non-recoverable antecedents.

Table 2a. Absolute value of Distance Marking with value of P and D

Corpus	Proximal	Distal
Plain	890	625
Dialogue	643	227

In table 2a the total number of feature "distance marking" of pronoun are 2385; out of which 890 are proximal and 625 of distal for plain corpus. In dialogue, proximal are 643 and distal 227. Inference from the above numbers is that both the values are important for anaphora resolution. Because the count of proximal and distal values is significant in number. The feature set "distance marking" is important and can make a significant impact on the analysis.

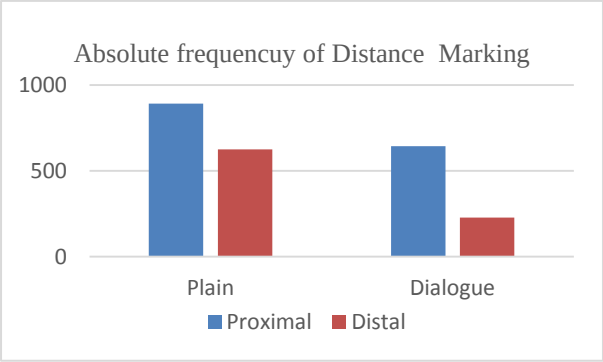

Fig. 1. Absolute frequency of Distance Marking

Table 2b. Percentage of Distance Marking with value of P and D

Corpus	Proximal	Distal
Plain	58.7	41.3
Dialogue	73.9	26.1

In figure number 2, the percentage of feature is kept on the Y axis and value of Distance Marking on X axis, which is P and D. P has Fifty eight (58.7) percentage of pronouns in plain text and forty one (41.3) percentage of "distal". It means 58.7 and 41.3 of pronoun are "proximal" and "distal" from its antecedent in plain text and considerable count for study of anaphora resolution.

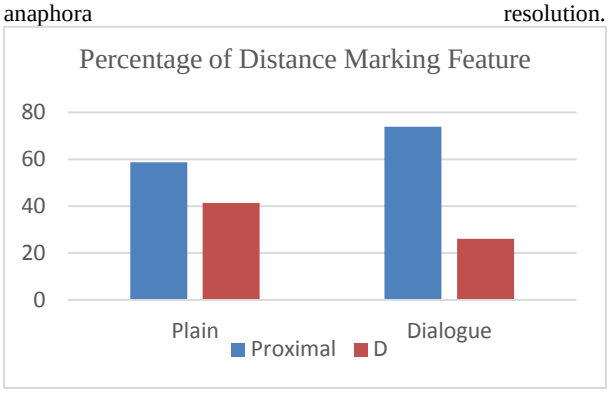

Fig. 2. Percentage of pronoun feature Distance Marking with value of P and D

On the other hand, these values in dialogue text are 73.9 and 26.1 with P and D value. The number of antecedents as pronoun in dialogue are more as compare to "distal" value in plain text. It implies that number of distal antecedent are less in number in plain genre. However, it is reversed in case of the first value of the Distance Making feature (P) in plain text and dialogue. Percentage of P value of pronouns in dialogue is 73.9 and for plain text 58.7. The inference of the discussion is proximal value in both texts is more as compared to and dialogue text and both have impact on anaphora resolution. The percentage and absolute value of distance marking have a substantial amount in the corpus. So, it is important to consider and discuss this feature in anaphora resolution.

4.2 Nature of Deixis

Table 3a.Total Count Nature of Deixis with value of P and D

Corpus	Pronoun	Demonstrative
Plain	585	930
Dialogue	197	674

The feature 'nature of deixis' has only two valid values P (pronoun) and D (demonstrative) third value is zero. In plain text there are 585 pronouns in the corpus and 930 demonstratives. Dialogue have 197 and 674 respectively, P and D. It means demonstrative are dominant in any corpus. But pronoun value is also considerable.

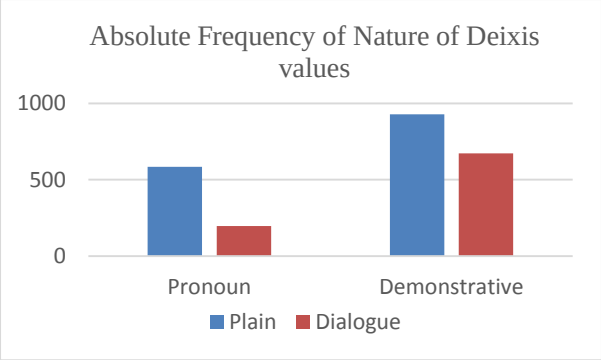

Fig. 3. Absolute frequency of Nature of Deixis

Table 3b. Percentage of Nature of Deixis with value of P and D

Corpus	Pronoun	Demonstrative
Plain	38.6	61.4
Dialogue	22.6	77.4

In figure number 3 percentage of feature kept on the Y axis and the value of 'Nature of Deixis' on X axis, which is P and D. Value of both P and D in above feature in plain and

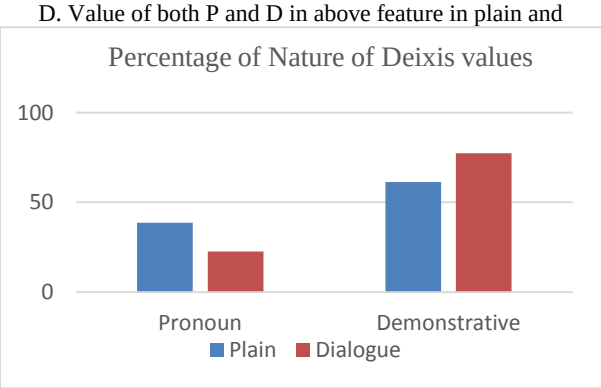

Fig. 4. Percentage of pronoun feature 'Nature of Deixis' with value of P and D

dialogue shows that the pronoun count is relativly less (38.6 & 22.6 in plain and dialogue respectively) as compared to demonstrative in in both texts, which is 61.4 & 77.4 in plain and dialogue respectively. Demonstratives are larger in number in both texts for "nature of deixis" feature. The percentage and the absolute value of P and D, both required to be addressed for anaphora resolution.

5. CONCLUSION

The given features and their respective percentage shows the genre behavior. These observations have been made;

- Maximum antecedents are noun phrases in the context of the demonstrative pronouns.
- Anaphora is much higher than cataphora. It is because anaphora is about 88 percentage (approximately) and the remaining 12% is cataphora.
- Almost 90 percentage antecedents are directly recoverable. Only 6 and 4 percentages (approximate) respectively are indirect anaphora in plain and dialogue.
- The "nature of deixis" shows that the majority of antecedents are demonstratives[2] (61 and 77 percentage) plain and dialogue corpus respectively and the rest are pronouns.
- There is a difference in all features and their values for plain data and dialogue.
- The above result may vary with change of the corpus. However, the basic features and their behavior will remain unchanged significantly.

Further, a machine can be trained to implement any artificial intelligence technique to resolve co-reference.

6. END NOTES

1. Final Annotation scheme for Hindi proposed by S.Botley [17] and Shiraj Sinha [18] and available at http://www.lancaster.ac. uk/fass/projects/corpus/emille/Hindi_anaphora.htm (accessed on 11/08/2014).
2. Demonstratives are considered as determiner. They are not considered as pronoun by some researchers. However reference [16] classify into two categories. i.e. Distal demonstratives, Modifier demonstratives.

7. REFERENCES

[1] Gambhir, V., "Syntactic restrictions and discourse functions of word order in standard Hindi," Doctoral Dissertation, Univ. of Pennsylvania, Philadelphia, Penn (1981).

[2] Prasad, R., Strube, M., "Discourse Salience and Pronoun Resolution in Hindi," In Penn Working Papers in Linguistics, 6.3. UPenn pp. 189-208 (2000).

[3] Botley, S. P., "Indirect anaphora: Testing the limits of corpus-based linguistics," International Journal of Corpus Linguistics, 11(1), pp 73–112, 2006.

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

[5] Dutta, K., Kaushik, S., Prakash, N., "Machine Learning Approach for the Classification of Demonstrative Pronouns for Indirect Anaphora in Hindi News Items," The Prague Bulletin of Mathematical Linguistics No. 95, pp 33–50, doi: 10.2478/v10108-011-0003-4, (2011).

[6] Prasaad, R., Miltaski, E., Joshi, A., Webber, B., "Annotation and Data Mining of the Penn Discourse Tree Bank," In ACL Workshop on Discourse Annotation, (2004).

[7] Hammami, S., Belguith, L. H., Hamadou A. B., "Arabic anaphora resolution: corpora annotation with coreferential links," In The International Arab Journal of Information Technology - IAJIT , vol. 6, no. 5, pp 480-488, (2009).

[8] Sinha, S., "A Corpus-based Account of Anaphor Resolution in Hindi," Master's thesis, University of Lancaster, UK, (2002).

[9] Singh, P., Dutta, K., "Sentence Structure for Free Word Order Language in Context with Anaphora Resolution: A Case Study of Hindi," International Conference on Computer Design Engineering and Technology (ICCDET 2014), vol:8 no:6 part XIX, pp 2011-2014, June 29-30, 2014 at London, United Kingdom.

[10] Singh P., Dutta, K., "Analysis and Comparison of Antecedent Type of Demonstrative pronoun in Context of Co-reference Resolution: A Corpus Based Study of Hindi for Monologue and Dialogue," Sixth IEEE International Conference on Computational Intelligence and Communication Networks (CICN 2014), pp 536-540, 14-16 Nov. 2014, DOI 10.1109/.122 537 DOI 10.1109/CICN.2014.122

[11] Singh P., Dutta, K., "Semiautomatic annotation scheme for demonstrative pronoun considering indirect anaphora for Hindi", IEEE symposium of NLP of International Conference on Advances in Computing, Communications and Informatics (ICACCI, 2014)," pp 1710 - 1714, India, 24-27 Sept. 2014, Print ISBN: 978-1-4799-3078-4, DOI:10.1109/ ICACCI. 2014. 6968538.

[12] Swift, M., Allen, J., and Gildea, D., (2004), "Skeletons in the parser: using a shallow parser to improve deep parsing," In Proceedings of the 20th international conference on Computational Linguistics (COLING '04). Association for Computational Linguistics, Stroudsburg, PA, USA, , Article 383 . DOI=10.3115/1220355.1220410 http://dx.doi.org/10.3115/1220355.1220410

[13] Esteve, Y., Bazillon, T., Antoine, J. Y., Béchet, F., Farinas, J.,"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, Valletta, Malta, may 2010. ELRA.

[14] Hinrichs, E., Zastrow, T. "Automatic Annotation and Manual Evaluation of the Diachronic German Corpus TüBa-D/DC," Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pp. 22-29. 2012.

[15] Palmer, M., Gildea, D., Kingsbury, P., "The Proposition Bank: An Annotated Corpus of Semantic Role," Computational Linguistics archive, Vol 31, Issue 1, pp. 71-106, March 2005.

[16] Botley, S., Mcenery, T., "Proximal and Distal Demonstratives A Corpus-Based Study," Journal of English Linguistics, vol 29; pp 214-233, 2001, DOI: 10.1177/00754240122005341

[17] Botley, S., 2000, "Corpora and Discourse anaphora: using corpus evidence to test theoretical claims," Ph.D. thesis, Lancaster University.

[18] Sinha, S., 2003, "Demonstrative anaphors in Hindi newspaper reportage: a corpus-based study" MA dissertation, Lancaster University.