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

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.
the link between the antecedent and anaphor 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 semi-automatic/automatic annotation [11-15]. Word order imposes more constraints [1], [9]. The five features 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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value1</th>
<th>Value2</th>
<th>Value3</th>
<th>Value4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance marking</td>
<td>P(proximal)</td>
<td>D(distal)</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Nature of deixis</td>
<td>P (Pronoun)</td>
<td>D (Demonstrative)</td>
<td>Z (Zero)</td>
<td>None</td>
</tr>
</tbody>
</table>

Example 6:
एक समाल के जवाब में मी सी.वी.आर्बर के अभ्यस्त वीडियो वेबसाइट ने रखा किया कि आभियुक्त के लिए विभिन्न विषयों का वाक्य बनाने के बाद सी.वी.आर्बर, अभ्यस्त शीर्ष की होती थी।<w tag="ne,s">में</w> <o tag="PDDARMN">>प्रसाद</o> ने कहा कि वह अपने नवश्वास अपने शक्तिशाली न्यायिकी बिना बेरोजगारी बनाने के लिए चलाया।</w>

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 EMILLE 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.

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

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Proximal</th>
<th>Distal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plain</td>
<td>890</td>
<td>625</td>
</tr>
<tr>
<td>Dialogue</td>
<td>643</td>
<td>227</td>
</tr>
</tbody>
</table>

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.

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.

4.2 Nature of Deixis

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.
The “nature of deixis” shows that the majority of antecedents are demonstratives (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


2. Demonstratives are considered as pronoun by some researchers. However reference [16] classify into two categories. i.e. Distal demonstratives, Modifier demonstratives.

7. REFERENCES


