Application of Pronominal Divergence and Anaphora Resolution in English-Hindi Machine Translation

Kamlesh Dutta, Nupur Prakash, and Saroj Kaushik

***Abstract*—So far the majority of Machine Translation (MT) research has focused on translation at the level of individual sentences. For sentence level translation, Machine Translation has addressed various divergence issues for large variety of languages; the issue of pronominal divergence has been presented only recently. Since the quality of translation as required by users follows coherent multi-sentence discourse structure in a specific context, the pronominal divergence helps us in understanding the nuances of translation arising out of disparity in the languages. Subsequently using clues from this divergence, the anaphora resolution system can find the correct interpretation for the given pronominal referents and other entities by resolving the inter-sentential context. In the literature, researchers have examined the issue and have proposed ways for their classification and resolution of anaphora. However for Indic languages, not many studies are available. In this paper, we discuss different aspects of pronominal divergence that affects the anaphora resolution in English Hindi Machine Translation (EHMT). The study shall be helpful in developing approaches that can explicitly use inter-sentential information in order to resolve specific types of ambiguity and which can generate coherent multi-sentence discourse structure in the target language to produce higher quality of translation Machine Translation.**

***Index Terms*—Pronominal, anaphora, machine translation, divergence.**

I. Introduction

THE syntactic, semantic and discourse level divergence in natural languages poses difficulty in the translation within two languages. Most of the machine translation systems have tried to capture the syntactic and semantic divergence as the translation takes place at the sentence level. The progress at the level of discourse is still at its infancy stage as it requires multi sentence level translation. One of the most important aspects in successfully analyzing multisentential texts is the capacity to establish the anaphoric references to preceding discourse entities. The paper will discuss the issue of pronominal divergence between two languages and the problem of anaphora resolution from the perspective of EHMT. The study shall be helpful in developing approaches that can explicitly use inter-sentential information in order to resolve specific types of ambiguity and which can generate coherent multi-sentence discourse structure in the target language to produce higher quality of translation MT.

Pronominal divergence between English and Hindi is expressed by the variation in the representation, e.g., English phrase “*It* is raining” has a corresponding translation as “*baarish ho rahi he*” (lit. “*rain is happening*”) in Hindi. Though typically, “*it*” has a corresponding translation as “*yeh*” or “*veh*”, in the given example “*it*” would have no mapping. For a native speaker or for an expert human translator, this may be a simple and obvious choice, the frequent occurrence of such divergence poses difficulty for the machine translation system. For example a good machine translation will be able to detect that “it” maps to “*veh*” or “*yeh*” in most of the cases, but it will be unable to detect the cases where the translation of “*it*” has to be dropped. Preliminary investigation on a sample text reveals that the divergence of this type is prevalent. Thus finding a way to deal with such a divergence shall help not only in the correct anaphoric resolution but also help in the quality translation.

In the literature ([1], [2], [3]), researchers have examined the issue and have proposed ways for their classification and resolution of anaphora. However for Indic languages, not many studies are available. In this paper we discuss different aspects of pronominal divergence that affect the anaphora resolution in English-Hindi Machine Translation (EHMT). We take classification of pronominal divergence approaches adopted by Mitkov in [2] and Gupta and Chaterjee in [4] as a starting point for our study about pronominal divergence and anaphora resolution in the translation of English and Hindi.

Once we are able to deal with the pronominal divergence between two languages, we shall be not only able to find the correct anaphoric references in the text but shall be able to generate the correct translation for the same. Section II presents the case of pronominal divergence between English and Hindi. Section III presents how pronominal divergence can be used in anaphora resolution. Section IV presents how machine translation systems can benefit from anaphora resolution. Finally, we conclude in section V with the future scope and the difficulties in employing anaphora resolution system for Hindi.

Manuscript received March 23, 2008. Manuscript accepted for publication March 04, 2009.

Kamlesh Dutta is with Computer Science & Engineering Department, National Institute of Technology, Hamirpur-177005 (HP), India (phone: +91-1972-3044424; fax: +91-1972-223834, e-mail: kdnith@gmail.com).

Nupur Prakash is with School of Information Technology, Guru Gobind Singh Inderprastha University, Delhi. Currently she is on deputation as additional director, ICAI, India (e-mail: nupurprakash@rediffmail.com).

Saroj Kaushik is with Computer Science & Engineering Department, Indian Institute of Technology. Delhi, India (e-mail: saroj@cse.iitd.ac.in).

II. Pronominal Divergence in EHMT

Pronominal divergence in EHMT as proposed by Gupta and Chatterjee in [4] pertains to the usage of *"it"*. Four types of the identified pronominal divergence are as follows:

1. Conversion of subjective compliment in English sentence into subject in the corresponding translation.
2. Conversion of adjectival compliment of the subject into subject.
3. Conversion of infinitive verb into subject.
4. Conversion of main verb into subject.
5. No divergence if *"it"* is a subject.

To illustrate these cases, let us have a look at the examples from Gupta and Chatterjee [4].

1) a) *"It is morning."*
subaha ho gayii hai
morning become has
b) *"It was a dark night."*
ek andherii raat thii
one dark night was

2) *"It* is *very humid today."*
aaj bahut umas hai
today very humidity is

3) *"It is difficult to run in the Sun."*
dhoop mein daudhnaa kathin hai .
Sun-shine in to run difficult is

4) *"It is raining."*
barsaat ho rahii hai.
rain be ing is

5) *"It is crying."*
veh ro raha/rahi hai.
He/she cry …ing is

The pronominal divergence as shown for *"it"* reveals that if the subject of the English sentence is not *"it"*, or if the subject of the Hindi sentence is *"veh"* or *"yeh"* then pronominal divergence will not take place. However, depending upon the subjective compliment or main verb of the English sentence the type of the pronominal divergence can be identified.

III. Anaphoric Properties of *"it"*

The pronominal divergence discussed in Section II can handle only single sentence translation. Incorporating anaphora resolution component in machine translation enables us to handle the discourse correctly by enabling multisentential translation. From anaphoric point of view the pronominal divergence cases are actually the subset of anaphoric references. From anaphoric point of view *"it"* can have following anaphoric properties as classified by Evan in [5] (examples are taken from this work).

(i) Nominal Anaphoric

"Do not sweep the dust$_i$ when dry, you will only recirculate it$_i$."

Pronoun *"it"* refers to nominal expression *"the dust"*.

(ii) Clause Anaphoric,

"One day in 1970, fifty thousand women marched down Fifth Avenue in New York. It$_i$ is said to have been the biggest women's gathering since suffrage days."

Pronoun *"it"* refers to the preceding clause in the text.

(iii) Proaction

"Mays walloped four home runs in a span of nine innings. Incidentally, only two did it$_i$ before a home audience."

Here *"it"* along with *do* refers to the preceding verb phrase.

(iv) Cataphoric

"When it$_i$ fell, the glass$_i$ broke".

The pronoun is coreferential with the next nominal expression in the text.

(v) Discourse Topic

"Always use a tool for the job it was designed to do. Always use tools correctly. If it$_i$ feels very awkward, stop."

The interpretation of the pronoun depends upon the context in which the pronoun is used.

(vi) Pleonastic

"It is worth having more than one size or a good-quality set with interchangeable bits."

In this case no interpretation for the pronoun.

(vii) Idiomatic/stereotypic,

"I take it you're going now."

The pronoun is non-referential, but used in certain fixed expressions in the language.

TABLE I
Anaphora and Pronominal Divergence

Anaphora	**Translation of *"it"* in Hindi**	**Divergence**
Nominal Anaphora	*us-ko/use*	Case-based
Clausal Anaphora	*yeh*	Case-based
Proaction	*us-ko/use*	Case-based
Cataphoric	*veh*	Case-based
Discourse Topic	-	Pronominal
Pleonastic	-	Pronominal
Idiomatic	-	Pronominal

Cases (i)-(iii) are anaphoric, which is to say that for a given pronoun an antecedent exist in the preceding text. Case (iv) suggests a forward search strategy. No explicit interpretation is available for the remaining cases. The translation of pronoun *"it"* occurring in each example (i)-(vii) in Hindi shows different translations (Table I). Case (i) and (iii) *"veh"* takes the accusative form and hence is inflected for *us-ko/use*. Case (ii) and (iv) takes the ergative form and hence the case divergence occurs in these examples. Examples shown in (v)-(vii) fall in the category of pronominal divergence.

IV. Anaphoric Reference and Divergence in EHMT

The discussion presented in section III shows anaphoric properties of *"it"* and we observe that the corresponding translation of "it" in Hindi is not similar. So is the case with other pronouns. Different anaphoric categories impose the constraints on the translation. The ambiguity in the translation can be resolved by incorporating syntactic, semantic or discourse related knowledge about the pronoun. Consider for example the following sentence:

6) "*The boys ate the sweet because they were hungry*."

A translation word-by-word into Hindi would require specifying correct case marking for "*The boys*" (for ergative case - *ne*) and would require assigning correct gender information to the verb phrase in the subordinate clause depending on the association of pronoun with its antecedent. The pronoun *"they"* can be translated as *"ve"* either of the form (third person, male, plural; third person, female, plural) reflected in the auxiliary verb, depending on the gender of its antecedent. Giving a random or default translation is not an option in this case, since it can lead to a target text with incorrect meaning. In order to generate the correct Hindi pronoun along with correct verb phrase, we need to be able to identify the correct antecedent of the English pronoun "*they*", which is *"the boys"*. If the antecedent is identified incorrectly as being *"the sweets"*, the error propagates into the Hindi translation, which becomes:

7) "*ladakon ne mithaiyan khaeen kyunki ve bhookhhi theen*."

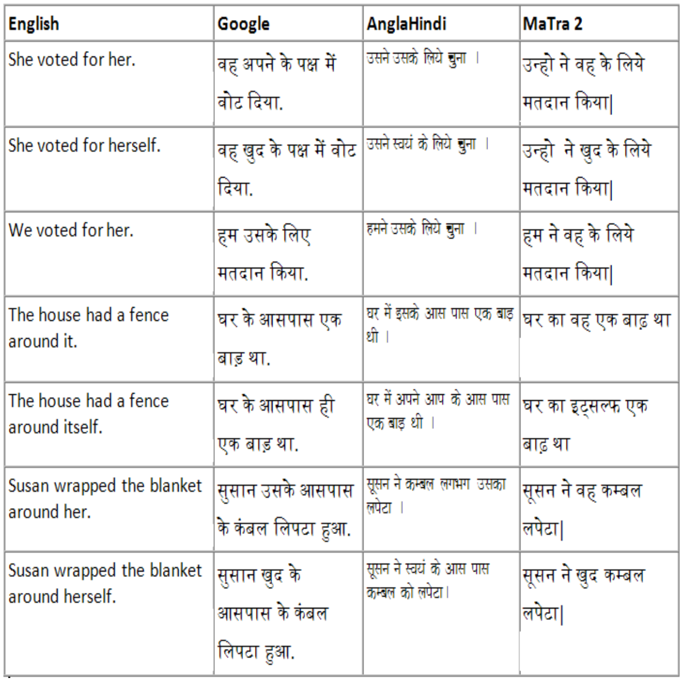
In this sentence, the pronoun *"ve"* can only be interpreted as referring to *"sweets" (*since this is the only possible antecedent that agrees in gender with the pronoun), therefore the message conveyed is "*The boys ate the sweets because the sweets were hungry"*, which is obviously not the intended meaning.

As is evident from the above example, the inherent divergence between the language pair poses certain difficulties. The interpretation of pronouns is made more difficult by the fact that pronouns offer very little information about themselves. All they convey is some morphological and syntactical information, such as number, gender, person and case. These considerations justify the interest that researchers showed towards developing systematic approaches for anaphora resolution (and in particular for pronominal anaphora) in naturally occurring texts. Incorrect translation of anaphoric relation in Hindi could be attributed to the following facts:

– Gender of pronouns from one language does not have a corresponding gender translation in another language,
– Language pairs have gender discrepancy,
– Distinction between animate and inanimate antecedents occurs,
– The indirect speech sentences in Hindi and English differ in both forms of tense and the use of pronominal elements
– Significant role played by case system,
– Other morphological features such as association of gender information with the verb clause in Hindi.

To substantiate our justification for the need of anaphora resolution in Machine translation, we translate English sentences into Hindi (Table II) using "AnglaHindi" [6], "MaTra2" [7] and Google service [8]. The corresponding English interpretation of translated sentences is tabulated in Table III. The evaluation for anaphora resolution of all these systems shows that apart from other issues as discussed by Dorr in [9] and Dorr et al in [10]; pronominal translation is affected by the lack of anaphora resolution in the system. Google translation is not able to resolve the ambiguity between nominative and ergative forms of subject pronouns. The verbal association fails to take into account the importance of auxiliary verb. The gender association with inanimate objects is ambiguous. MaTra2 fails to specify correct form of pronouns occurring in the object position. Further it fails to translate "itself" and "ourselves" as well. Even the gender association is incorrect in few sentences as evident from Tables II and III. Anglahindi, on the other hand is better than the other two translation systems. The system has problem in making a choice of correct reflexive pronouns.

TABLE II
Translation of Pronominal Sentences



English	Google	AnglaHindi	MaTra 2
She voted for her.	वह अपने के पक्ष में वोट दिया.	उसने उसके लिये चुना ।	उन्हो ने वह के लिये मतदान किया|
She voted for herself.	वह खुद के पक्ष में वोट दिया.	उसने स्वयं के लिये चुना ।	उन्हो ने खुद के लिये मतदान किया|
We voted for her.	हम उसके लिए मतदान किया.	हमने उसके लिये चुना ।	हम ने वह के लिये मतदान किया|
The house had a fence around it.	घर के आसपास एक बाड़ था.	घर में इसके आस पास एक बाड़ थी ।	घर का वह एक बाढ़ था
The house had a fence around itself.	घर के आसपास ही एक बाड़ था.	घर में अपने आप के आस पास एक बाड़ थी ।	घर का इट्सल्फ एक बाढ़ था
Susan wrapped the blanket around her.	सुसान उसके आसपास के कंबल लिपटा हुआ.	सूसन ने कम्बल लगभग उसका लपेटा ।	सूसन ने वह कम्बल लपेटा|
Susan wrapped the blanket around herself.	सुसान खुद के आसपास के कंबल लिपटा हुआ.	सूसन ने स्वयं के आस पास कम्बल को लपेटा।	सूसन ने खुद कम्बल लपेटा|

TABLE III
CORRESPONDING INTERPRETATION OF TRANSLATED SENTENCES

English	**Google**	**AnglaHindi**	**MaTra2**
She voted for her.	*He voted for himself*	*He/She selected for him/her*	*They voted for he/she*
She voted for herself.	*He voted for himself*	*He/She selected for himself/herself.*	*They voted for themselves*
We voted for her.	*We voted for him/her*	*We selected for him/her*	*We voted for he/she*
The house had a fence around it.	*The house had a fence around it*	*In the house, it had a fence around her.*	*This was a fence of the house*
The house had a fence around itself.	*Around the house only, there was a fence.*	*In the house, around itself, there was a fence.*	*The house had its own fence.*
Susan wrapped the blanket around her.	*Susan her around blanket wrapped around her*	*Susan blanket approximately her wrapped.*	*Susan wrapped that blanket.*
Susan wrapped the blanket around her.	*Susan of around herself blanket wrapped*	*Susan wrapped around herself blanket.*	*Susan wrapped blanket herself.*

V. CONCLUSION

Pronominal divergence can help in identifying anaphoric and non-anaphoric occurrences of pronoun. Case based divergence helps us in identifying the correct inflection form for the corresponding pronoun for EHMT. Our studies of *"it"* pronouns reveals that the pronominal divergence is a subset of anaphoric classification. Since majority of Machine Translation systems only handle one-sentence input, the use of pronominal divergence has limited application for MT. For the further improvement in the translation, processing of multiple sentences for resolving the correct antecedent and thereby generating the correct anaphor (pronoun) is much more useful. Perhaps looking at the complexity involved in understanding and incorporating anaphora resolution majority of the machine translation systems preserve anaphora ambiguities to be corrected by user latter on. Still, the challenge involved in the problem has not deterred the researcher. With the amount of research being conducted in the area of anaphora resolution since last decade, one can be optimistic to have quality automated translation work in the near future.

REFERENCES

[1] R. Mitkov, *Anaphora Resolution*, Pearson Education. Longman, London. 2002.

[2] R. Mitkov, S. K. Choi and R. Sharp, "Anaphora Resolution in Machine Translation," in *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation TMI 95*, pp. 87-95, Leuven, Belgium, 1995.

[3] A. F. Gelbukh and G. Sidorov, "On Indirect Anaphora Resolution," in *Proc. PACLING-99, Pacific Association for Computational Linguistics*, pp. 181-190, Waterloo, Ontario, Canada, August 25-28, 1999.

[4] D. Gupta and N. Chaterjee, "Identification of Divergence for English to Hindi EBMT," in *Proceeding of MT Summit- IX*, pp. 141-148, 2003.

[5] R. Evans, "Applying Machine Learning Toward an Automatic Classification of It," *Literary and Linguistic Computing,* Vol. 16. No. 1, Oxford University Press, pp. 45-57, 2001.

[6] http://www.cse.iitk.ac.in

[7] http://202.141.152.9/matra/index.jsp

[8] http://translate.google.com/

[9] B.J. Dorr, "Machine Translation Divergences: A Formal Description and Proposed Solution," *Computational Linguistics*, Vol. 20, Number 4, pp. 597-633, 1994.

[10] B. J. Dorr, L. Pearl, R. Hwa and N. Habash, "DUSTer: A Method for Unraveling Cross-Language Divergences for Statistical Word-Level Alignment," *Machine Translation: From Research to Real Users, LNCS 2499*, pp. 31-43, 2003.