USING RICH MORPHOLOGY IN RESOLVING CERTAIN HINDI-ENGLISH MACHINE TRANSLATION DIVERGENCE

R. Mahesh K. Sinha

Indian Institute of Techmology, Kanpur 208016 India
rmk@iitk.ac.in

Abstract

Identification and resolution of translation divergence (TD) is very crucial for any automated machine translation (MT) system. Although this problem has received attention of a number of MT developers, devising general strategies is hard to achieve. Solution to the language specific pairs appears to be comparatively tractable. In this paper, we present a technique that exploits the rich morphology of Hindi to identify the nature of certain divergence patterns and then invoke methods to handle the related translation divergence in Hindi to English machine translation. We have considered TDs encountered in Hindi copula sentences and those arising out of certain gaps in verb morphology.

Introduction

Translation Divergence (TD) occurs when the underlying concept of a sentence gets manifested differently in different languages. For instance, in Hindi causative verb is expressed morphologically whereas in English it is structurally expressed (*raam ne kapaRe dhu**lavaayaa*** {Ram ERG clothes wash-CAU} => Ram **got** clothes **washed**). The resolution of such TDs is very crucial for any machine translation system. In general, these tasks are difficult to accomplish and TD remains one of the significant research topics in machine translation (MT) area. The topic has been examined from different perspectives in the literature on MT (Dorr 1994, Habash et al 2002) and different approaches have been proposed both for their classification and their resolution.

These studies have revealed the complexity of the problem. Although this problem has received attention of a number of MT developers, devising general strategies is hard to achieve. On the other hand, solution to the language specific pairs appears to be comparatively tractable. Here we have considered Hindi-English language pair for our investigation. The translation divergence among English and Hindi has been studied in Sachi et al (2001), Gupta et al (2003) and Sinha et al (2005). In this paper, we have taken the TD classification and details as given in Sinha et al (2005) which is the latest and fairly detailed on the subject. We present a technique that exploits the rich morphology of Hindi to identify the nature of some of these divergence patterns and then invoke methods to handle the related translation divergence in Hindi to English machine translation. The rich morphology of Hindi (Kachru 1980) provides a lot of cues about the surrounding words and the sentence structure without creating a complete parse of the sentence. The rich morphology has been exploited in part-of-speech tagging for Hindi (Gupta et al 2006).

In this work, we have considered TDs encountered in Hindi copula sentences and TDs arising out of certain gaps in verb morphology (Sinha et al 2005). The following sections present details of the technique used.

Handling Divergence in Copula Sentences

In Hindi possession, location and existential (English 'be' form) constructs are expressed using copula verbs (linking verbs such as *hai, haiN, thaa, the, ho etc).* The examples (1-7) illustrate this aspect of Hindi grammar.

(1) *usakaa eka betaa hai* (possession).
(His one son is)
'He has one son'

(2) *raama usakaa betaa hai.*(existential)
(Ram his son is)
'Ram is his son'

(3) *raama kaa eka shikshaka thaa.*(possession)
(Ram of one teacher was)
'Ram had a teacher.'

(4) *raama ke paasa eka kalama hai* (possession)
(Ram of near one pen)
'Ram has a pen'.

(5) *dilli ke paasa aagaraa hai.*(location)
(Agra Delhi of near is)
'Agra is near Delhi'.

(6) *shera jangala meN hai.*(location)
(Lion forest in is)
'The lion is in the forest.'

(7) *raama eka shikshaka thaa.*(existential)
(Ram one teacher was)
'Ram was a teacher.'

It is observed that all the example (1)-(6) , use some post-positions ('Kaaraka' based 'Vibhakti' marker Sinha (1989)). The post-position used in examples (1)-(3) is *'kaa'*, in examples (4)-(5) *'ke paasa'* and in example (6) it is *'meN'*. These post-positions denote morphological attributes and are easily extractable. We use these

morphological attributes in association with neighbouring words to detect the nature of translation divergence. For example, if the post-position *'kaa/ke/kii'* is found with only one noun-phrase preceding it, it is recognized as the possessive ('has/have/had' construct in English) and if it is preceded by two noun-phrases, it is existential ('be' construct in English). Where there is no post-position as in example (7), it is taken as existential.

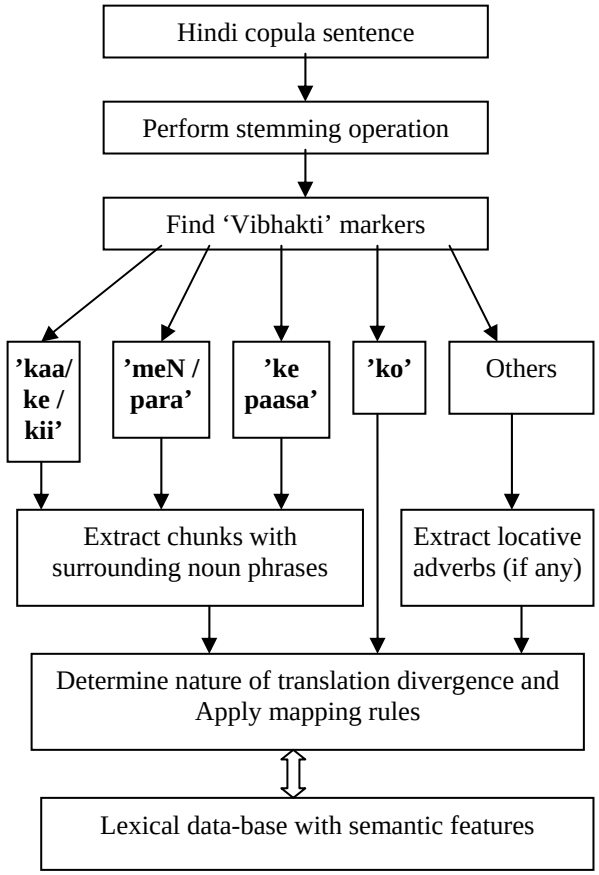

Figure 1: Handling Translation Divergence in Hindi Copula Sentences

The figure 1 presents block schematic depicting our strategy at a broad level. When the MT system detects the input sentence to be a copula sentence, this module is invoked. The stemming operation extracts the root words in the sentence. For the post-positions *'kaa/ke/kii, meN /para, ke paasa'*, chunks are formed with surrounding noun phrases based on the nature of the post-position. These chunks along with morphological attributes in the form of post-positions and the semantic features of the words are directly used in determining whether the copula sentence is possessive, locative or existential. In case of post-position *'ko'*, no chunking is performed. In all other cases (for any other 'Vibhakti' marker or no marker) , locative adverbial phrase, if present, is fed to the next block. Following sections describe the process for each of these cases.

'kaa/ke/kii': 'Sambandha Kaaraka'

Following are some examples with *'kaa/ke/kii'* post-positions. Analysis for each one these is given along with it.

8. *usake chaara bete haiN.*
(His four sons are)
'He has four sons'
Post-position: '*ke'*
Chunk formed: *usake chaara bete*
Rule used: Sambandh Kaarak chunk followed by copula verb is possessive.

9. *usake chaara bete bahaadura haiN.*
(His four sons brave are)
'His four sons are brave'
Post-position: '*ke'*
Chunk formed: *usake chaara bete*
Rule used: Sambandh Kaarak chunk followed by adverbial phrase before copula verb is existential.

10. *ye usake chaara bete haiN.*
(These his four sons are)
'These are his four sons'
Post-position: '*ke'*
Chunk formed: *usake chaara bete*
Rule used: Sambandh Kaarak chunk preceded with a noun phrase and followed by copula verb is existential.

11.*raama kaa eka shikshaka hai.*
(Ram of one teacher is)
'Ram has a teacher.'
Post-position: '*kaa'*
Chunk formed: *raama kaa eka shikshaka*
Rule used: Sambandh Kaarak chunk followed by copula verb is possessive.

12. *raama kaa eka ghara do manjilaa hai.*
(Ram of one house two storied is)
'Ram's one house is double storied'
Post-position: '*kaa'*
Chunk formed: *raama kaa eka ghara*
Rule used: Sambandh Kaarak chunk followed by adverbial phrase before copula verb is existential.

'meN/para': 'Adhikarana Kaaraka'

The divergence as pointed out in examples (13) pertaines to 'there' construction in English. This is a very difficult divergence to detect. Further, Hindi does not use determiner (corresponding to 'the' and 'a' in English) in an explicit manner. However, the notion of the determiner is contained in the word order of the sentence. The rules used in examples (13) and (14) demonstrate how easily our strategy works not only the 'there' construct but also the determiner.

13. *jangala meN eka shera hai.*
(Forest in one lion is)
'There is a lion in the forest.'
Post-position: '*meN'*
Chunk formed: *jangala meN eka shera*

Rule used: Adhikaran Kaarak chunk followed by copula verb is locative indefinite and English construct is of 'there' type.

14. *shera jangala meN hai.*
(Lion forest in is)
'The lion is in the forest.'
Post-position: '*meN'*
Chunk formed: *jangala meN*
Rule used: Adhikaran Kaarak chunk preceded by a noun phrase and followed by copula verb is locative definite and for English 'the' be used as determiner for the common noun phrase which becomes the subject.

For the examples (15) and (16) , the analysis is same as those for (13) and (14) respectively except that the post-position here is *'para.*
'
15. peRa para totaa hai.
(tree on parrot is)
'There is a parrot on the tree.'

16. *totaa peRa para hai.*
(parrot tree on is)
'The parot is on the tree'.

It is interesting to note that in the examples (17) and (18), a single change of morphological attribute (*usakaa* to *usake*) completely changes the meaning of the sentence and our strategy is able to capture this. It does give wrong result if inflectional form is same as the non-inflectional form.

17. *usakaa ghara para durghatanaa huaa.*
(his house on accident be PST)
'He met with an accident at home'
Post-position: '*para'*
Chunk formed: *ghara para* (It is important to note here that *usakaa ghara para* is not allowed to become a single chunk as the post-post *para* demands that the preceding noun phrase chunk is in oblique form (see example 18).
Rule used: Adhikaran Kaarak chunk preceded by a Sambandh Kaarak chunk and with copula verb is locative with noun in Sambandh Kaarak chunk becoming the subject.

18. *usake ghara para durghatanaa huii.*
(his house on accident be PST)
'There was an accident at his house.'
Post-position: '*para'*
Chunk formed: *usake ghara para* (see comments in example 17 above).
Rule used: Adhikaran Kaarak chunk with copula verb is locative and English construct is of 'there' type.

'ke paasa' : Locative/Possessive cases

19. *dilli ke paasa aagaraa hai.*
(Agra Delhi of near is)
'Agra is near Delhi'.
Post-position: '*ke paasa'*
Chunk formed: *dilli ke paasa*
Rule used: *ke paasa* chunk with copula verb and another noun phrase of semantic type 'place', is locative.

20. *usake paasa kuttaa hai.*
(dog his near is)
'He has a dog.'
Post-position: '*ke paasa'*
Chunk formed: *usake paasa*
Rule used: *ke paasa* chunk with its noun of semantic type 'animate' with copula verb preceded with another noun phrase, is possessive with 'have/has/had' construct in English and the other common noun is indefinite.

21. *kuttaa usake paasa hai.*
(dog his near is)
'The dog is with him.'
Post-position: '*ke paasa'*
Chunk formed: *usake paasa*
Rule used: *ke paasa* chunk with its noun of semantic type 'animate' preceded with another noun phrase, and with copula verb is possessive with 'with' construct in English and the other common noun is definite.

'ko' : Dative cases

22. *raama ko bukhaar hai.*
(Ram to fever is)
'Ram has fever'.
Post-position: *'ko'*
Rule used: It is possessive with other noun.

23. *raama ko jaanaa hai*
(Ram to go is)
'Ram has to go'.
Post-position: *'ko'*
Rule used: If the copula verb is preceded with verb-root form ('*naa/ne/nii'* ending), it is of 'has/have/had'+'to <verb>' form in English.

Others

24. *baahar bahuta thanda hai.*
(outside very cold is)
'It is very cold outside)
Post-position: none
Locative adverb: *baahar*
Rule used: if the copula verb is preced with a noun with semantic type of 'weather', the English construct is of the type 'it'.
It should be noted that another diverge pattern which is hard to capture.

25. *raama imaanadaara hai.*
(Ram honest is)
'Ram is honest.'
Post-position: none
Locative adverb: none
Rule: Default existentential.

Handling Certain Main Verb-based Translation Divergence in Hindi

The basic strategy followed here to perform morphological analysis of verb chunk and other morphological attributes in the sentence. These morphological attributes are then matched with the divergence pattern of Hindi. If the match is found then the corresponding English pattern is generated as per the transformation rules designed to deal with the divergence pattern. Figure 2 outlines this basic strategy.

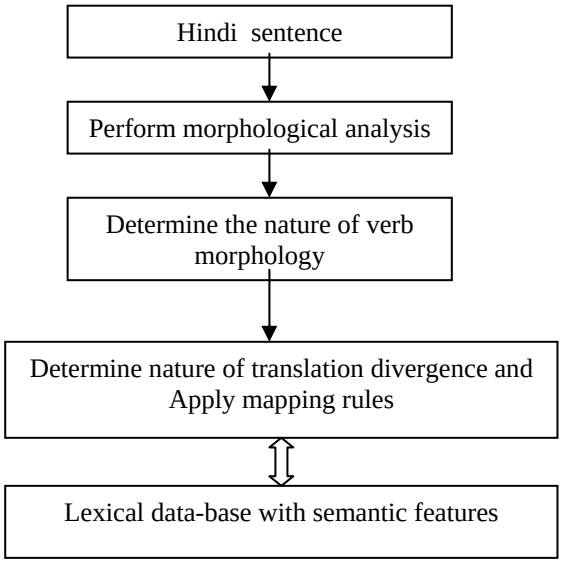

Figure 2: Handling Certain Main Verb-based Translation Divergence in Hindi

An analysis of the classes of verb-based divergence patterns (Sinha et al 2005) considered here is elaborated in the following subsections.

Impersonal Passive Constructions

(26) *raama se chalaa nahin jaataa*.
{Ram by walk not go PASS}
'Ram cannot walk.'

When the verb is morphologically ending with *'taa/te/tii'* with no auxiliaries, and the post-position *'se'* is used with the subject, the corresponding English pattern is 'can <verb>'.

Habitual Aspect

(27) *vah yahaaN aayaa karataa thaa.*
{he here come.HAB do.IMP}
'He used to come here.'

When the verb is morphologically ending with *'yaa/ye/yii'* followed by *'karataa/karate/karatii'* + *'thaa'*, the corresponding English pattern is 'used to <verb>'.

(28) *vah yahaaN aayaa karataa hE.*
{he here (often) come.HAB do.IMP be.PR}
'He comes here.'

When the verb is morphologically ending with *'yaa/ye/yii'* followed by *'karataa/karate/karatii'* + *'hE'*, the corresponding English pattern is '<verb> in present tense form'.

(29) *vah bolataa rahaa.*
{he speak.IMP PROG}
'He kept on speaking.'

When the verb is morphologically ending with *'taa/te/tii'* followed by *'rahaa/rahe/rahii'*, the corresponding English pattern is 'kept on <verb> in present participle form'.

Certain Transitive and Causative verbs

The transitive and causative forms of most of the intransitive verbs in Hindi are derivable through morphological rules that are easily programmable. An example is an intransitive verb *'haNsanaa'* (laugh), is morphologically transformed to *'haNsaanaa'* (transitive:make laugh) and *'haNsavaanaa'* (causative: get make laugh). Similarly causative verbs are derived with morphological transformation on the transitive verb. An example is the transitive verb *'khariidanaa'* (buy), is morphologically transformed to *'khariidavaanaa'* (causative: get bought). Once the pattern is identified, corresponding English pattern is generated in English using light verbs 'make' and 'get'.

(30) *raama haNsaa.* [Intransitive]
{Ram laughed}
'Ram laughed.'

(31) *raama-ne siitaa-ko haNsaayaa.* [Transitive]
{Ram-ERG Sita-ACC laugh-TRS}
'Ram made Sita laugh.'

(32) *raama-ne siitaa-ko mohan se haNsavaayaa.* [Causative]
{Ram-ERG Sita-ACC Mohan-by laugh-CAUS}
'Ram got Mohan make Sita laugh.'

Optative Sentences

These sentences are composed of two clauses. The main clause uses 'wish' verb (wish, want, desire, pray etc) and the verb form in the subordinate clause morphological end with *'ao/oN/eN/uN'* based on the person of the subject.

(33) *ham caahate hEN ki aap saphal hoN.*
{we want be.PR that you successful be.OPT}

'We want that you succeed.'
Or, 'We want you to succeed.'

Let-sentences

The verb in these type of sentences is morphologically ending with *'ne'* + *'do/deN/diijjiye'*.

(34) *use jaane do*
{him go give IMPR}
'Let him go.'

There is another form of 'Let' sentence that is used with first person plural form. Given below is an example:

(35) a. *aao, hamaloga kheleN.*
{come, we play IMPR}
'Come on, let us play.'

b. *chalo, nadii meN taireN.*
{go, river in swim IMPR}
'Come on, let us swim.'

Here the verb morphologically ends with *'eN'* and the sentence is imperative. The sentence starts with words like *'aao/chalo/ hey'* denoting some kind of addressing or command. Further, the usage of first person word *'ham/hamaloga'* is optional and is implied (as in example (35 b) above).

Had-Counterfactual Clause

This kind a 'had' construct is very peculiar in English. The corresponding pattern in Hindi has two clauses joined with an conjunction *'to'* (then) and one clause starts with *'agar/yadi'* (if) The verb in both the clauses morphologically end with *'taa/te/tii'*. Once the pattern is identified, corresponding English pattern is generated in English using the third form of the verb in the 'had' clause and 'would have' in the second clause.

(36) *agar tum yahaaN hote to ham bhii aate.*
{if you here be.SUBJ then we also come-SUBJ}
'Had you been here we would have also come.'

Conclusions

The above examples and their analysis amply demonstrate the fact that the morphological attributes of a morphologically rich Hindi language provide a number of cues that are helpful in identifying a number of translation divergence patterns in Hindi to English machine translation. The methodology does not require input sentences to be completely parsed and no part of speech tagging is performed. These rules have been tested with about 1000 random sentences taken from the corpus of Hindi stories available at the web-site:
http://ildc.gov.in/hindi/hdlbooks.htm.
The accuracy achieved in case of copula Hindi sentences was about 85% and for the main verb-based classes it was about 60%.

The rules have been hand-crafted. However, these rules can be acquired using a Hindi to English parallel corpus. It is interesting to note that these rules yield generalized patterns for example-based machine translation.

Abbreviations:
ACC: Accusative Case, AFF: Affirmative, CAUS: Causative, CONT: Continuative Aspect, CPP: Conjunctive Participial Particle, DAT: Dative Case, DET: Determiner, DIT: Ditransitive, DUR: Durative Aspect; ERG: Ergative Case, EW: Echo Word, FU: Future Tense, GER: Gerund, HAB: Habitual Aspect, IMP: Imperfective Aspect, IMPR: Imperative Mood, INT: Interrogative, OPT: Optative Mood, PASS: Passive Particle, PR: Present Tense, PST: Past Tense, QP: Question Particle, RP: Relative Pronoun, SUBJ: Subjunctive Mood, TRS: Transitive, VPRT: Verbal Participle.

References

Dorr, Bonny (1994). Classification of Machine Translation Divergences and a Proposed Solution. Computational Linguistics, 4(20), 597-633.

Sachi, D., J. Parekh and P. Bhattacharya (2001) Interlingua-based English-Hindi Machine Translation and Language Divergence. Machine Translation 16 (4),:251-304.

Gupta Deepa, and Niladri Chattejee (2003) Identification of Divergence for English to Hindi EBMT. In Proceeding of MT Summit-IX: 141-148.

Gupta, Kuhoo, Manish Shrivastava, Smriti Singh and Pushpak Bhattacharyya (2006) "Morphological Richness Offsets Resource Poverty- an Experience in Building a POS Tagger for Hindi", COLING/ACL, Sydney, Australia.

Habash, N. and Bonnie Dorr (2002). Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation. LAMP-TR-088, CS-TR-4369, UMIACS-TR-2002-49.

Kachru, Yamuna (1980) Aspects of Hindi Syntax. Manohar.Delhi.

Sinha, R.M.K. (1989) "A Sanskrit based Word-expert model for machine translation among Indian languages, Proc. of workshop on Computer Processing of Asian Languages," Asian Institute of Technology, Bangkok, Thailand: 82-91.

Sinha R.M.K. and A. Thakur. (2005). "Divergence Patterns in Machine Translation between Hindi and English". In Proceeding of MT Summit X, Phunket, Thailand: 346-353.

http://ildc.gov.in/hindi/hdlbooks.htm