Divergence Patterns in Machine Translation between Hindi and English

R. Mahesh K. Sinha
Indian Institute of Technology Kanpur
rmk@iitk.ac.in

Anil Thakur
Indian Institute of Technology Kanpur
anilt@iitk.ac.in

Abstract

The issue of translation divergence is an important research topic in the area of machine translation. An exhaustive study of the divergence issues in MT is necessary for their proper classification and resolution. In the literature on MT, scholars have examined the issue and have proposed ways for their classification and resolution (Dorr 1993, 1994). However, the topic still needs further exploration to identify different sources of translation divergence in different pairs of translation languages. In this paper, we discuss translation patterns between Hindi and English of different types of constructions with a view to identifying the potential topics of the translation divergences. We take Dorr's (1993, 1994) classification of translation divergence as the base to examine the different topics of translation divergence in Hindi and English. The primary goal of the paper is to point out different types of translation divergences in Hindi and English MT that have not been discussed in the existing literature.

1 Introduction

The issue of translation divergence is a complex topic in machine translation (MT). The translation divergence can be defined in terms of language-to-language differences in the respective grammars. Thus a divergence occurs when a sentence in language L_1 translates into a sentence in L_2 in a very different form[1] (Dorr 1994: 12). The topic has been studied from different perspectives and a number of approaches have been proposed to handle them. It is crucial for any MT system to identify the nature of translation divergences and resolve them so as to obtain correct translation. The translation divergences occur at different levels and affect the quality of the translation according to the degree of complexity involved in a particular translation divergence. It has also been noted (Dorr, 1994) that certain types of translation divergences are universal in the sense that they exist across the languages whereas certain other types of translation divergences are specific to a pair of translation languages. Therefore, the translation divergences need to be studied from both across-language and language-specific perspectives. In this paper, we examine Hindi and English translation language pair largely from the perspective of identifying the language-specific divergecnes. Hindi and English differ in many respects and hence this translation language pair[2] presents a rich source for the study of translation divergence in MT. These languages also show significant differences from the point of view of socio-cultural perspectives that need to be properly examined. In this paper, we discuss different aspects of Hindi and English grammars that involve potential areas of translation divergences in Hindi and English MT. We discuss divergence issues in Hindi-English machine translation and the same translation pair is then examined for reverse translation from English-Hindi so as to examine the nature of the divergence in each case.

In the existing literature, the issue of translation divergence for Hindi and English MT has not been exhaustively examined. Gupta et al (2003) and Dave et al (2001) discuss some of the translation divergences pertaining to English-Hindi MT and Hindi-English MT. Dave et al (2001) discusses the issue within the UNL-based Interlingua approach and only some of the obvious types of divergences have been discussed. These works do not explore further areas of divergence. Some of the obvious divergence types such as thematic divergence, dative divergence and movement divergence have not been discussed at all. Although the authors point out divergences resulting from the pro-drop phenomenon in Hindi and the occurrence of pleonastic subjects in English, they do not examine the issue in detail to capture the implications of

[1] It should be noted that what constitutes a translation divergence is not dependent upon the translation strategy used for machine translation. This is contrary to the views expressed by one of the reviewers. In this paper we have taken this definition of divergence and presented structural differences both in forward and reverse directions irrespective of the MT strategy.

[2] One of the reviewers has correctly pointed out that languages other than Hindi which are equally distant from English can be assumed to exhibit similar or more such divergences as discussed in this paper. Natural languages are very complex and no research on translation divergence can be said to be exhaustive, particularly at this stage of research.

these language-specific features for other types of divergences. Also, some of the examples that have been discussed under head-swapping divergence such as promotional and demotional divergences need to be re-looked for their proper categorization. For instance, *on* (as in "the play is on" => *khel cal rahaa hE* {play go PROG be.PR} has been taken as an adverbial element in English which has a verbal realization in Hindi. However, if we recognize this use of '*on'* (meaning in Hindi as '*caalu'*) as an adjectival element, the divergence no longer exists. The Hindi translation (*khel caalu hE* {play on be.PR} for the English sentence ("the play is on.") is equally valid and a commonly used sentence. Gupta et al (2003) discusses only a few cases of divergence to present rules for unification of translation divergences in English-Hindi MT. Thus we notice that the existing works are far from exhaustive both from the point of view of classification and resolution of different translation divergences in the context of Hindi-English MT.

In section 2, we discuss different sources of translation divergences in Hindi and English MT. Section 3 presents a brief outline of strategy used in dealing with these divergences in our MT system followed by the concluding remarks.

2 Translation Divergence: Classification and Further Issues

Dorr (1993) categorizes translation divergences into two broad types. They are: (A) Syntactic Divergences, (B) Lexical-semantic Divergences. They are further subcategorized as follows:

(A) Syntactic Divergence: i. Constituent order divergence, ii. Adjunction divergence, iii. Preposition-stranding divergence, iv. Movement divergence, v. Null subject divergence, vi. Dative divergence, and vii. Pleonastic divergence

(B) Lexical-semantic Divergence: i. Thematic divergence, ii. Promotional divergence, iii. Demotional divergence, iv. Structural divergence v. Conflational divergence, vi. Categorial divergence, and vii. Lexical divergence

In Dorr (1994), she has examined the structure of the lexical-semantic divergences and proposed a LCS-based approach for their resolution. This classification takes into account various sources of differences between a set of translation language and captures a large sets of translation divergences. The classification is based on the Government and Binding framework (Chomsky 1986, Jackendoff 1990) of linguistic theory which assumes a deep structure to capture the surface structure variations. The deep structure functions as the universal structure, i.e. applicable across languages. Thus both the classification and the resolution of the translation divergences are largely discussed from the perspective of the universal grammar. The classification captures the major grammatical issues in translation divergence across languages. However, it also misses a number of points that pertain to a particular set of translation languages. The issue of divergence between a set of languages is associated with a number of factors ranging from linguistic to socio- and psycho-linguistic aspects of the languages involved. Although Dorr's classification takes into account many of the major linguistic factors associated with translation divergence, there still remains a number of points related to both linguistic and extra-linguistic factors that may exist in different sets of translation languages. Furthermore the parameters of the classification does not take into account subtle semantic factors to the extent they are relevant for the classification of translation divergences in various languages. Without going into a detailed discussion of the different classes and categories of translation divergences as proposed in Dorr (1993, 1994), we discuss English and Hindi translation examples that present new sources and topics of translation divergence in English-Hindi and Hindi-English MT.

2.1 Non-Configurational Nature of Hindi

English is a configurational language that follows a rigid word order pattern as opposed to Hindi which is relatively less rigid and exhibit free word order variation. This is one of the major sources of divergence between a pair of natural languages. In Dorr's classification, word order related translation divergences have been discussed under syntactic divergence. Dave el al (2001) extends Dorr's classification to English-Hindi translation pair but do not discuss the implications of the word order facts at all. For instance, one of the implications of the word order related divergence can be noticed with respect to the interpretation of the question particle *'kyaa'* (Sinha et al. 2005c) in Hindi. '*kyaa'* can be used both as a marker of interrogative pronoun in content question sentences and as a question particle in yes-no question sentences. Besides certain other factors such as the category of the verb, it is the position of occurrence of '*kyaa'* that indicates its interpretation one way or the other. The particle '*kyaa'* in the sentence-initial and sentence-final positions are generally interpreted as question particle rather than as an interrogative pronoun, as is evident from the examples shown in (1).

(1) a. *aap **kyaa** paDh rahe hEN?* {you what read PROG be.PR} => **What** are you reading?
b. ***kyaa** aap paDh rahe hEN?* {QP you read PROG be.PR} => **Are** you reading?

The examples in (1)[3] show subtle implications with respect to the word order facts in Hindi. Replicative and Echo Words

Hindi, like most of the other South Asian languages, exhibits the phenomenon of replication (Sinha et al. 2005d) of the lexical items to express different grammatical functions. The English counterparts of these Hindi constructions do not resort to replicative structure. This distinction often results into a change in the syntactic category of the relevant elements. For instance, we notice that in Hindi, as in (2), the replication of the verb (in participial form) denote an adverbial function of cause. The English counterpart of this function is realized by a gerundive prepositional phrase.

(2) *vah **bolate bolate** thak gayaa.* {he speak speak tired got} => He got tired **of speaking**.

In this example, the replicative element *bolate bolate* is an adverbial clause which is realized lexically in Hindi and is mapped in English structurally. The reverse translation for this example set does not involve divergence[4], as in (3).

(3) He got tired **of speaking**. => *vah **bolane se** thak gayaa.* {he speak of tired got}

Another typological feature exhibited by all the Indian languages is the occurrence of echo words where a lexical word is partially replicated to denote a wide range of meanings with subtle semantic constraints. The examples in (4-5) are illustrative.

(4) ***caay vaay** pii kar jaaiye.* {tea EW drink CPP go} => Have some **snacks** before going.

(5) *ise Thiik se **jaaNc vaaNc** lo.* {this properly examine EW take} => Please **examine** it properly. <=> *ise Thiik se jaaNc liijiye*.

The echo words generally have no lexical status in the lexicon of the language. However, whenever an echo word is identical with a lexical word, it affects the interpretation of the preceding lexicon. In (4), the use of an echo word '*vaay'* along with the main word '*caay* (tea) gives the sense of light refreshment. However, this is not a possible sense in which an echo word is used in (5). Here the main verb *jaNcanaa* 'examine' occurs with an echo word that has only an emphatic (or extension) function but it cannot be exactly expressed in English. In the case of the reverse translation from English to Hindi no divergence is encountered.

2.2 Expressive Elements

Expressive words exist in all natural languages and pose difficulty in processing, particularly in mapping onto another language. The reason is that these words do not have exact parallel in another language. Thus the word *dhaRaam* is only distantly mapped by '*bump*' in English, as in (6).

(6) *vah **dhaRaam** se girii.* {she 'dhaRaam' with fell} => She fell with a 'bump'.

The expressive words usually originate from the sound associated with the semantics of the action verb and can be adverbial or verbalized action-verbs such as '*tap-tapaanaa*' (drip), '*khat-khataanaa*' (knock) etc. One may argue this to be just a lexical gap but indeed it is not so. However, some of these words can be handled in the lexicon but as in many cases the mapping also involves structural changes, the issue involves a wider scope of interpretation.

2.3 Asymmetry in NP and Existential Clauses

The issue of divergence related to the difference in the determiner systems of English and Hindi NPs has not been examined in the existing literature on divergence. English has (in)definite articles that mark the (in)definiteness of the noun phrase overtly whereas Hindi lacks an overt article system and different devices are used to realize the (in)definiteness of a noun phrase in Hindi. For instance, mapping onto articles *a-an/the* in English is not lexically realizable from Hindi (e.g. *laRakaa aayaa* => The/*a boy came.). In this connection, another point of divergence between Hindi and English related to *there-* and *it*-sentences in English is worth examining. In English, *there-* and *it*-constructions are used to denote existential sentences (besides others). Hindi does not have a pleonastic subject construction and the contrast between existential and non-existential (mostly definite) sentences is realized by several other ways such as the movement of the noun phrase from its canonical position and the use of demonstrative elements. Let us look at the examples in (7-8).

(7) *kamare meN **saaNp** hE.* {room in snake be.PR} => **There** is a snake in the room.

(8) ***saaNp** kamare meN hE.* {snake room in be.PR} =>**The snake** is in the room.

We notice that the bare noun phrase *saaNp* 'snake' in (7) and (8) is mapped by indefinite and definite noun phrases in English. However, the only difference between these two Hindi sentences is the respective positions of the subject NP and the (place) adverbial phrase. When we look at the

[3] **ACC:**Accusative Case, **AFF:**Affirmative, **CAUS:**Causative, **CONT:**Continuative Aspect, **CPP:**Conjunctive Participial Particle, **DAT:**Dative Case, **DIT:**Ditransitive, **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, **SUBJ:**Subjunctive Mood, **TRS:** Transitive, **VPRT:**Verbal Participle.

[4] In case of multiple possible translations, if any one the translations exhibit the same grammatical structure, it is considered as a case of no divergence.

reverse translation of the same translation sentence, the nature of divergence is different.

(9) There is a snake in the room. => *kamare meN ek saaNp hE.* {room in a snake be.PR}

Hindi does not have a counterpart of "there-construction" and the Hindi grammar has to resort to a number of devices such as shifting of the relevant elements and deletion of '*there*' to obtain the equivalent of the English sentence, as in (9).

2.4 Tense, Moods and Aspects (TAM)

Another important source of translation divergence in Hindi and English MT is associated with the difference in the manifestation of different tense, moods and aspectual properties of the verb in these languages. For instance, Hindi uses a certain type of passive construction that marks a kind of non-volition function. The English counterparts of such Hindi sentences are only partially able to express the exact meaning.

(10) ***raam se*** *galatii ho gaii.* {Ram by mistake be PASS} => **Ram** made a mistake. <=> *raam-ne galatii kii.*

The possible English counterpart of the Hindi sentence in (10) is far from the actual sense in which the Hindi impersonal passive has been used. The literal sense will be somewhat like: 'a mistake got made by Ram unintentionally'. Thus the reverse translation for the same translation sentence from English to Hindi involves far more complex procedure [5]. A somewhat similar dimension of divergence between Hindi and English is manifested with respect to the negative impersonal passive constructions in Hindi and the way they are realized in English.

(11) ***raam se*** *calaa nahiiN jaataa.* {Ram by walk not PASS} => **Ram** cannot walk. <=> *raam cal nahiiN sakataa.*

In this case, too, no translation divergence occurs in the case of the reverse translation and the source Hindi sentence cannot be obtained.

In Hindi, some of the aspectual features of the verb are realized by verbal inflection whereas English resorts to different non-inflectional ways such as phrasal verb or an adverbial element or a prepositional phrase with gerund as the head, to realize them. For instance, in (12-13), the aspectual property is identical in both the sentences and the difference is located only in tense. The habitual aspect of the tense is reflected by inflectional morphology on the verb in both the tenses. However, this habitual aspect in English is realized by the use of a phrasal verb in the case of the past tense (12) and by the use of an adverbial word '*often*' in the case of the present (and future) tense (14). Thus the adverbial element in Hindi is optional whereas the one in English cannot be optional. In (14), we notice that the non-terminative aspect is realized by verbal morphology in Hindi whereas English uses a phrasal structure to realize this aspect.

(12) *raam* ***aayaa karataa*** *thaa*. {Ram come CONT be.PST} => Ram **used to** come.

(13) *raam* ***(aksar)*** *aayaa karataa hE*. {Ram often come CONT be.PR} => Ram ***(often)** comes.

(14) *raam* ***bolataa rahaa***. {Ram speak CONT} => Ram **kept on speaking**.

In certain types of conditional clauses in Hindi, there is optainality between present and future/past tenses. But the English counterparts of these Hindi sentences always require the verb to occur in the present tense.

(15) *yadi tum dillii* ***jaate ho / jaoge*** *to tum kaamyaab hoge*.{if you Delhi go FU / PST then you successful be.FU} => If you **go** to Delhi you will be successful. <=> *yadi tum dillii* ***jaataa ho*** *to tum kaamyaab hogaa*. {if you Delhi go then you successful be.FU}

The reverse translation from English to Hindi will produce only the source Hindi sentence that has the verb in the present tense form and hence will not involve any translation divergence.

2.5 Role of Conjunctions and Particles

Another source of divergence between Hindi and English can be located in the case of the use of different conjunctions and particles in Hindi. We take examples involving some of these particles in Hindi such as *ki*, *na*, and *yaa*. The translation divergence between Hindi and English related to *ki* is quite complex (Sinha and Thakur, 2005b). *ki* is mainly used as a sentence complementizer, but can also be used to indicate alternate conjunction in an affirmative sentence (16) and an interrogative sentence (18) in Hindi.

(16) *siitaa mujhase milii* ***na ki*** *usase.* {Sita me met not him} => Sita met me **not** him.

(17) *raam paDhataa hE* ***ki*** *sotaa hE?* {Ram read PROG be.PR or sleep PRPG be.PR} => Does Ram study **or** sleep? <=> *kyaa raam paDhataa hE yaa sotaa hE*.

In another instance, *yaa* ('or') is a coordinate conjunction particle in Hindi that conjoins two clauses or phrases. However, it can denote a different function in Hindi depending on the punctuation mark used in the relevant sentence.

[5] One of the reviewers has argued that such a claim makes no sense as it can only be made in relation to a given system. The point we are making here is that it is not possible to derive a sentence to sentence translation whatever be the MT system. A translation can be only in the form of a number of sentences 'explaining' the situation.

For instance, when *yaa* ('or') is used in a sentence that has an interrogative marker, *yaa* functions as an interrogative marking particle. The contrast is shown in (18-19).

(18) *vah dilli gayaa hE **yaa** kolkata.* {he Delhi went be.PR or Kolkata} => He has gone either to Delhi **or** to Kolkata.

(19) *vah dilli gayaa hE **yaa** kolkata?* {he Delhi went be.PR or Kolkata}=> **Has** he gone to Delhi **or** Kolkata? => *kyaa vah dillii gayaa hE yaa kolkata?*

In the reverse translation of the English sentence in (19) back into Hindi, we notice that the use of an interrogative particle (*kyaa*) is obligatory.

2.6 Asymmetry in Transitivity and Causativity

The divergence related to the morphology-syntax asymmetry for Hindi-English translation pair can be located in the difference in the realization of certain transitive verbs and most of the causative constructions in Hindi and English.

(20) *raam-ne siitaa-ko **haNsaayaa**.* {Ram-ERG Sita-ACC make-laugh} => Ram **made** Sita **laugh**.

(21) *raam-ne siitaa-ko mohan se **haNsavaayaa**.* {Ram-ERG Sita-ACC Mohan-by make-laugh-CAUS} => Ram **got** Mohan **make** Sita **laugh**.

In Hindi, there are three forms of a verb (in this case *haNsanaa* 'laugh') that are morphologically derived. (*haNsanaa* => *haNsaanaa* => *haNsavaana*). The English counterparts of these sentences show that in English, there is only one lexical verb '*laugh*' and the other forms are realized by syntactic processes (such as resorting to different kinds of verbal constructions). In Hindi, *haNsaanaa* is a transitive verb which does not have a lexical counterpart in English (English has only the intransitive form as a lexical item). In English, it is realized by using two verbs *make* and *laugh. haNsavaanaa* is a lexical causative verb in Hindi which in English is realized by using three verbs *get*, *make* and *laugh*, with separate argument structures of their own. The English counterpart of the Hindi example in (21) appears to be a forced translation. In certain cases, it is quite difficult to obtain an exact translation of a common Hindi ditransitive verb. For instance, in (21), the English counterpart of the transitive verb *piinaa* is *drink*. However, Hindi also has a ditransitive verb *pilaanaa* derived from *piinaa.* English does not have a counterpart of this ditransitive verb.

(22) *usane hame paanii **pilaayaa**.* {he-ERG us water drink-CAUS} => he **gave** us water. <=> *usane hame paanii **diyaa**.*

The verb '*give*' is used because there is no exact English counterpart of the Hindi verb *pilaanaa*. Thus the reverse translation involves a far more complex procedure. Gaps of this type are quite common between Hindi and English.

2.7 Stative and Progressive Aspect

English seems to lack an exact counterpart of Hindi stative verb/adjective which is realized by the progressive aspect marker. In English, there is no distinction between the progressive aspect denoting sentence and its stative counterpart. The English verbs such as *sit*, *stand*, *sleep*, and *wake* fall in this category. In Hindi, they are distinguished by different lexical form of the relevant verb.

(23) *raam kursii par **bEThaa** hE.* {Ram chair on sitting be.PR} => Ram is **sitting** on a chair. <=> *raam kursii par **bEThaa rahaa hE**.* {Ram chair on sitting PROG be.PR}

The divergence of this kind seems to involve both lexical and structural aspects of the languages involved. If '*sitting*' is entered in the lexicon both as an adjective and a (form of) verb, only then the source Hindi sentence can be obtained in reverse translation.

2.8 Participle Modification

The participle modifiers in Hindi are mostly realized by relative clauses in English. For instance, in Hindi, *vaalaa* is a suffix that, besides denoting several other functions, also functions as an adjectival suffix.

(24) *kal aane-vaale logoN se mEN nahiiN mil paauNgaa*. {tomorrow coming-of people with I not meet able to} => I will not be able to meet the people who are coming tomorrow. <=> *mEN un logon se jo kal aa rahe hEN nahiiN mil paauNgaa.* {I those people with who tomorrow come PROG be.PR not meet able to}

In (24), we notice that in Hindi, the noun *log* ('people') is modified by a participial adjectival phrase *kal aane-vaale* that precedes the head noun. However, in English, the same is realized by a relative clause construction that follows the main clause. In this case, an (adjectival) phrase in Hindi is realized by a clause in English. The *vaalaa*-construction in Hindi presents a complex issue in itself which cannot be discussed in this paper.

2.9 Gerund and Participle Clauses

Another significant source of divergence in Hindi and English MT can be located in the way the various clausal complements and adjuncts (such as verbal participles) in Hindi are realized in English.

(26) *vah **aakar** khush huaa.* {he come-CPP happy be.PST} <=> He got happy **to come**.

(26) *vah mujhase baat **karane (ke liye)** aayaa.* {he me talk do for came} <=> He came **to talk** to me.

(27) *vah yah **karane meN** samarth nahiiN hE.* {he this do in able not be.PR} <=> He is not able **to do** this.

(28) *vah jaanaa caahataa hE.* {he go want} <=> He wants **to go**.

In the Hindi sentences in (25-28), the adjunct verbal clauses and complement verbal clauses are realized by different structures which in English are mapped by a single structure. Thus the reverse translation for this set of examples in (25-28) faces different type of difficulty. In the former case, it is many-to-one mapping whereas in the latter case, it is a one-to-many mapping.

2.10 Clausal Conjunction

Another difference that is manifested between Hindi and English is with respect to clausal conjunction where the subordinate clause is used to express different types of clause in Hindi. In English, they are not always realized by the same type of clause, rather they are realized by different devices such as a modal verb.

(29) ***ho naa ho** vah kahiiN gayaa ho.* {may be he somewhere gone SUBJ} => He might have gone somewhere. <=> *vah kahiiN gayaa hogaa.* {he somewhere gone be.FU}

As we notice, the reverse translation does not produce exactly the source Hindi sentence.

2.11 Mapping of *have-verbs* in Hindi

Certain English have-sentences are difficult to be exactly mapped onto Hindi. Besides its polysemous nature, *have*-constructions also involve structural aspects and constitute a case of translation divergence.

(30) a. *usameN saahas hE.* {he-in courage be.PR} <=> He has courage.

b. *usake tiin bacche hEN.* {he-of three kids be.PR} <=> He has three kids.

c. *usake paas pEse hE.* {he-of near money be.PR} <=> He has money.

Some of the representative examples, as in (30), can show the divergence issues involved in translating the Hindi counterparts of the English have-constructions. In Hindi, the subject NP occurs in different case forms but they all are mapped onto English by have-verb sentences. In the reverse translation from English to Hindi, although there will be one-to-many mapping but the nature of the divergence will remain the same.

2.12 Had-Counterfactual Clause

In Hindi, the counterfactual conditional clause is marked by a conjunction *agar/yadi* ('if') which in English can be realized either by a had-clause or an if-clause. In the former case, translation divergence occurs.

(31) *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.

In reverse translation for the English sentence, the divergence remains the same.

2.13 Let-sentences

The Hindi permissive-sentences are mostly translated into English by "let-sentences", as in (32a). However, there are certain wish-sentences that also occur in the form of a let-sentence (32b).

(32) a. *use jaane do*{him go give}<=>Let him go.

b. *calo, khaanaa khaayeN.* {go food eat.OPT} => Let us go and eat now. <=> *hameN jaane aur khaane do.* {us now eat let}

The difference in the English sentences between (32a) and (32b) is only in the use of a pronoun. The use of first person plural pronoun 'us' in (32b) makes the sentence a wish-sentence rather than a permissive-sentence. Thus the reverse translation in (32a) does not involve translation divergence whereas in (32b), a translation divergence occurs. The nature of this translation divergence again pertains to the gaps in the realization of different verbal inflections and functions between English and Hindi.

2.14 News Headings

The news headings in English and Hindi follow different grammar rules (Sinha 2002). In English, generally the present tense form of the verb is used whereas Hindi uses past tense from of the verb.

(33) *sunaamii meN laakhoN log **mare.*** {tsunami in millions died} => millions **die** in tsunami. <=> *sunaamii meN laakhoN log marate hE.* {tsunami in million people die}

In this case, translation in both the Hindi-English and English-Hindi cases involves divergence.

2.15 Optative Sentences

The verb in the subordinate clause in the sentences of optative mood in Hindi occurs in different forms depending on the gender, number and person of the subject NP of the subordinate clause whereas its English counterpart remains constant (root form of the verb) in all the cases. In this case, the verb form does not indicate tense. The divergence is triggered by the similarity of the form of the verb with other tense forms in the case of English whereas it is not the case in Hindi.

(34) *ham caahate hEN ki raam saphal ho*. {we want be.PR that Ram successful be.OPT} => We want that Ram succeed.

This type of divergence can be resolved by taking into account the semantic type of the verb.

2.16 Indirect Speech

The indirect speech sentences in Hindi and English differ in both the form of tense and the use of pronominal elements.

(35) a. *raam-ne kahaa ki mEN nahiiN jaauNgaa.* {Ram-ERG told that I not go-FU} => Ram said that I/he would not go. <=> *raam-ne kahaa ki mEN/vah nahiiN jaauNgaa / jaayegaa.* {Ram-ERG told that he not go-FU}

b. Ram said that I **was** coming. => *raam-ne kahaa ki mEN aa rahaa huuN.* {Ram-ERG said that I come PROG be.PR} => *Ram said that I **am** coming.

The use of the pronoun *mEN* 'I' in (35) is ambiguous and can be translated either by 'I' or 'he' in English. The example in (35b) shows that the tense in the English indirect speech sentences is past but must be mapped by present tense in the Hindi sentence. Although some aspects of this type of translation divergence have been partially discussed in Dave et al (2001) for Hindi-English MT, we notice that the issue needs further examination.

2.17 Socio-cultural Factors

Different natural languages have different mechanisms to indicate socio-cultural features leading to a variety of divergence in translation. We examine two of these in case of Hindi-English translation in the following sections.

2.17.1 Honorificity Markers

In Hindi, the honorific feature is marked by the pluralization of the verb and the use of plural pronominal elements whereas in English it is not the case.

(36) *raastrapati aa cuke hEN. ve ab bhaashan deNge.* {president come CPT be.PR. he now lecture deliver.FU}=> The President has arrived. He will deliver a lecture now. <=> *president aa cukaa hE. vah ab bhaashan degaa.* {president come CPT be.PR. he now lecture deliver.FU}

In both Hindi-English MT and English-Hindi MT, the divergence caused by this socio-cultural aspect of the respective language arises.

2.17.2 Mappings of Time

Usually, people's perception of different objects in the world is dependent upon several socio-cultural beliefs. For instance, time is conceptualized in the Indian culture differently than that is done in the Western culture. These concepts are expressed through our respective languages and difference in concepts manifests itself in the language that is the source of translation divergence. For instance, in English, the concept of a.m. vs. p.m. cannot be exactly mapped in Hindi. The Hindi counterpart of a.m. and p.m. denote only a small part of time and the other parts of time is denoted by different other terms. The example in (37a) shows that the time at the 5 o'clock in the morning is denoted by a.m. in English but the exact translation of a.m. in Hindi does not produce an appropriate Hindi expression. However, in (37b), we notice that the time at the 11 o'clock in the morning which is expressed in English by a.m. can also be expressed in Hindi by the exact translation of the term a.m. A similar situation is noticeable with respect to the mapping of p.m. in examples (37c-e).

(37) a. He arrived at 5 a.m. => *vah paaNc baje subah / *purvaahan ko aayaa.* {he five o'clock a.m. came} <=> He arrived at 5 o'clock in the morning.

b. He arrived at 11 a.m. <=> *vah gyaarah baje* **subah / purvaahan ko aayaa.* {he eleven o'clock a.m. at came}

c. He arrived at 3 p.m. <=> *vah tiin baje aparaahan ko aayaa.* {he three o'clock p.m. came}

d. He arrived at 5 p.m. => *vah paaNc baje shaam ko / *apraahan ko aayaa.* {he five o'clock evening at came} <=> He arrived at 5 o'clock in the evening.

e. He arrived at 7 p.m. => *vah sat baje raat ko / *aparaahan ko aayaa* {he seven o'clock night at came}. <=> He came at 7 o'clock in the night.

3 Dealing with Divergence in System Design

The foregoing discussions clearly show the existence of a number of issues that need detailed study to exhaust the empirical base for the classification of translation divergences in MT. It is also evident that in general, it is not possible to deal with all kinds of divergences in a machine translation system. For example, for translation from Hindi to English, certain TAM constructs and differentiating stative and progressive aspects of certain verbs, cannot be handled as English lacks a mechanism to represent the exact meaning. Barring these, we have dealt with most of the divergences in our machine-aided translation systems (Sinha 2004) for English to Hindi and Hindi to English translations. We have incorporated hybridization of rule-based and example-based strategies in our

systems. During the development stage, an attempt is made first to devise rules based on patterns of translation divergence in the rule-base. In case no definitive pattern emerges, such divergences are handled through example-base. The examples are stored in generalized form and distances computed with respect to syntactic and semantics tags associated with constituents. An input sentence is first matched with the example-base and on failure rule-based approach gets invoked. The English to Hindi translation system is primarily rule-based while the Hindi to English system is primarily example-based. We have also used paraphrasing to handle some of the divergences.

The divergences of the types under 'Non-Configurational Nature', 'Role of Conjunctions and Particles', 'Participle Modification', 'Gerund and Participle Clauses', 'Mapping of have-verbs in Hindi', 'Let-sentences', 'Indirect Speech' are handled through rules. All other types of divegences are dealt with example-based approach or through hybrid means combining paraphrasing, rules and examples.

4 Concluding Remarks

In this paper, we have examined the issue of translation divergence in Hindi-English MT keeping in view the classification of translation divergence as proposed in Dorr (1993, 1994) and some of the existing works on Hindi-English MT (Dave et al 2001, Gupta et al 2003, Sinha and Thakur 2005a). We have discussed mostly those translation divergences in English and Hindi MT that have not been discussed in the existing works. We have also discussed the reverse translations (English-Hindi) for some of the translation pairs to examine the nature of divergence (if any) in the case of reverse translation. We have observed that there are a number of areas in Hindi-English translation pair that fall under translation divergence but cannot be accounted for within the existing parameters of classification. We propose that to capture these (other) types of translation divergences from Hindi to English and vice-versa, we need to further modify the classification and augment it by new categories and subtypes. The topics of translation divergence discussed in this paper should provide insights into the complexity of translation divergences in Hindi and English and give a direction for their further classification and resolution.

Some of the earlier works on translation divergence between Hindi and English have been discussed to show their inaccuracies in terms of the coverage of the data and also the way some of the divergence issues have been approached. However, due to constraints on space, it has not been possible to include discussions on issues of divergence related to several particles in Hindi and mappings patterns of certain possessive constructions which in English are realized by 'have'-construction.

Further, due to space constraints, we have only briefly outlined strategies used our MT systems for translation from Hindi to English and vice-versa in dealing with these translation divergences.

References

S. Dave, J. Parikh and P. Bhattacharyaa. 2001. Interlingua-based English-Hindi Machine Translation and Language Divergence. *Machine Translation 16(4)*:251-304.

B. Dorr. 1993. *Machine Translation: a View from the Lexicon*. The MIT Press, Cambridge, Mass.

B. Dorr. 1994. Classification of Machine Translation Divergences and a Proposed Solution. *Computat. Linguistics 20(4)*:597-633.

B. Dorr, N. Ayan and N. Habash. 2004. Divergence Unraveling for Word Alignment of Parallel Corpora. *Natural Language Engineering*, *1(1)*: 1-17.

D. Gupta and N. Chatterjee. 2003. Identification of Divergence for English to Hindi EBMT. In *Proceeding of MT Summit-IX*: 141-148.

R.M.K. Sinha. 2002. Translating News Headings from English to Hindi, In *Proceedings of 6th IASTED International Conference on Artificial Intelligence and Soft Computing (ASC2002)*, Banff, Canada.

R.M.K. Sinha 2004. An Engineering Perspective of Machine Translation: Angla Bharti - II and Anu Bharti - II Architecture. In *Proceedings of International Symposium on Machine Translation NLP and TSS*, The McGraw-Hill Companies, New Delhi: 10-17.

R.M.K. Sinha and Anil Thakur. 2005a. Translation Divergence in English-Hindi MT. In the *Proceeding of EAMT Xth Annual Conference*, Budapest, Hungary, 30-31 May.

R.M.K. Sinha and Anil Thakur. 2005b. Handling *ki* in Hindi for Hindi-English MT. In the *Proceeding of MT Summit X*, Bangkok, 12-16 September.

R.M.K. Sinha and Anil Thakur. 2005c. Dealing with Replicative Words in Hindi for Machine Translation to English. In the *Proceeding of MT Summit X*, Bangkok, 12-16 September.

R.M.K. Sinha and Anil Thakur. 2005d. Disambiguation of 'kyaa' in Hindi for Hindi to English machine translation, *Sixth International Conference of South Asian Languages (ICOSAL-6),* Hyderabad, India, 6-8 January.