Deviation In Example Based Machine Translation – In Indian Perspective

Puran Krishen Koul
Asst Prof in Computer Science IILM Academy of Higher Learning

Abstract

The eminence of Example-Based Machine Translation (EBMT) which depends upon how efficient the modification scheme is. Adaptation essentially aims at modifying the retrieved examples to meet the required demands of a given translation system. One of the aspect of modification is handling with deviation. Here one looks at any structural difference that has to be incorporated due to inherent constraints of the source or target language. The present work looks at modification of for EBMT between English and Hindi. Special attention has given to the study of deviation by recognizing six different categories of deviation and providing schemes for identifying them.

1. Introduction

Example-Based Machine Translation (EBMT) [Nirenburg 1994] is based on the idea of performing translation by imitating translation examples of similar structures. In this type of translation system, a large amount of translation examples between two languages (L1 & L2, say) are stored in a textual database. These examples are subsequently used as guidance for future translation tasks. In order to translate a new input text in language L1, a similar L1 text is retrieved from the database, along with corresponding translated text in L2. This example is then adapted suitably to generate a translation of the input.

One major aspect of EBMT is its *modification* scheme. However good may be the similarity measurement scheme, and however large may be the textual database, in general there will not be an exact match for a given input sentence. Consequently, to carry out a translation task, adaptation has to be employed to modify the retrieved example to meet the current translation requirement. One major difficulty in adaptation is called "*deviation*". In general, deviation occurs due to some inherent incompatibility between the source and target languages. Study of adaptation therefore needs a careful study of deviation too.

In this paper, we discuss the issue of adaptation, in general, with special emphasis on deviation. Section 2 discusses different types of adaptation schemes that may be applied for English-Hindi machine translation. Section 3 deals with the concept of deviation in detail. Section 4 discusses some basics pertaining to deviation identification. Section 5 provides techniques for identifying different types of deviation between English and Hindi

2. Adaption of Translated text

Upon retrieval, an EBMT scheme looks for generating a translation for the input sentence with the help of the retrieved example(s). In general, this means consideration of the discrepancy between the input sentence and retrieved sentence in L1 first. The retrieved L2 sentence is then modified with the help of these discrepancies.

Five broad schemes for adaptation may be identified:

2.1. Simple word replacement or deletion

One can get the translation of the input sentence by replacing some words in the retrieved translation example. Suppose the input sentence is: "*Puran is eating rice*."

The most similar sentence retrieved by the system (along with its Hindi translation) is: "*Rinku is drinking water*." (*rinku paanii pii rahaa hai).*

In order to generate the translation, one just needs to replace "*Puran*" by "*Rinku*", "*eat*" by "*drink*" and "*rice*" by "*water*". Therefore, only word replacement gives the exact translation of the input sentence. In some cases one may have to delete some words from the translation example to generate the new translation. For example, the input sentence is: "*Rinku has given the book.*" The retrieved translation example along with its Hindi translation is" *Rinku has given the book to Mary"* (*rinku mary ko kithab dai chuka hai.*)

The translation can then be obtained by deleting the “*to Mary (mary ko)*” part from the retrieved translation.

2.2. Word addition

Sometimes to generate a new translation one may have to add some additional words in the retrieved translation example. For illustration, one may consider the example given just above with the roles of input and retrieved sentences being reversed.

2.3. Judicious word replacement

If the input and the retrieved sentence have some common words that have different equivalents in the L2 language, suitable replacement of some word may be needed.For example, suppose the input is:“ *She is taking tea.*” The corresponding retrieved example is: “She is taking rice. (wah chaawal khaa rahi hai)” Although both the sentences have same verb “take” their Hindi equivalents are different “khaanaa” for “rice” and “piinaa” for “tea”. So the verb has to be chosen judiciously.

2.4. Change in tense

when the input and retrieved sentences are different in the tense, one has to apply syntax rules for appropriate modification of the retrieved translation example. If it is “is + verbIstform +ing”implies “*rhaa hai*”, “verbIstform + s/es” implies “*taa hai*” or “*tii hai*”. For example if the input sentence is: “*Puran is eating rice*.(*Puran haawal khaa rahaa hai*)” The retrieved sentence is: “*Puran eats rice*. (*Puran chaawal khaata hai*)”

Then for generating the translation one has to replace “*khaata*” by “*khaa rahaa*” to adhere to the grammar rules.

2.5. Deviation

Special structural difference between the sentences, which we discuss below.

The first four items can be accomplished by studying the syntactic and semantic properties of the languages and forming appropriate rules. Some such general structural properties of English and Hindi are described in [Rao, 1998]. For example, The basic sentence pattern in English is Subject (S) Verb (V) Object (O), whereas it is SOV in Hindi. Consider for example “*Sita saw Gita”* here “Sita” is subject; “*saw*” is the verb while “*Gita*” is the object. So the words occur in the order SVO. But in Hindi it becomes “*sita ne gita ko dekha*” (SOV). English is a positional language, and is therefore (relatively) fixed-order. Hindi is (relatively) free-order. For illustration, “Ram killed Ravana” is very different from “Ravana killed Ram” but in Hindi “r*aam ne raavana ko maaraa*” has the same meaning as “*raavana ko raam ne maaraa*”.

In English, the modifiers of an object can occur both before and after the object. For example, adjectives usually precede nouns, whereas preposition phrases usually follow noun. In Hindi, modifiers usually occur before the object they modify. For example: “*The bay of Bengal”* is translated as *“bangal kii khaadii”* in Hindi. Such structural differences need to be identified for different contexts and appropriate rules have to be formed.

However, all translation adaptations are not very systematic. For example, “*Puran walks slowly”* can be written in Hindi as “*Puran dhiire se chalta hi*”. Thus the adverb “*slowly*” gets mapped into the adverbial phrase “*dhiire se*”. But similar modification cannot be made for “*Ram eats hungrily*”. The correct Hindi of this is “*Puran bhukho kii tarah khaataa hei*” (Puran eats like a hungry person). This is because in Hindi there is no suitable adverbial phrase to represent the adverb “*hungrily*”. Such discrepancies in representation are primarily due to some inherent characteristics of the languages (both source and target). Such difference in representation between two languages is called “*deviation*”. The existence of translation deviations makes the straightforward transfer from source structures into target structures difficult [Dorr, 1994].

Deviation can be of two broad categories:
2.5.1. Syntactic deviation
2.5.2. Lexical – Semantic deviation

The difference between these two types of deviations is that the former category is characterized by syntactic properties associated with each languages (i.e., properties that are independent of the actual lexical items that are used) whereas the later category is characterized by properties that are entirely lexically determined. In this work we concentrated on the second type of deviation. In Section 3., will focus on deviation of lexical-semantic type.

3. Deviation in English Hindi Translation

The issue of deviation originates in the following way. Corresponding to each language we have some mechanism for realizing the semantics of a sentence from its syntax. using lexical-semantic knowledge. Suppose we consider a sentence in the source language L1, and its semantics is realized using the source language related knowledge. Now we consider the corresponding translation in L2. If there is any difference in the roles of its different constituents with respect to the semantics of the L1 sentence, then deviation arises.

Evidently, deviation is language-to-language phenomenon. Dorr’s work is based on English-Spanish and English-German translations. Based on these two language pairs 7 different categories have been identified. However, we have not so far found suitable examples for all the 7 types of deviations in English-Hindi translation. But we have identified some other deviations between English and Hindi that are not found in Dorr's work.. Below we present examples of different types of deviations that we have discovered having gone through different parallel texts:

3.1. Thematic Deviation

The verbal object in one language becomes as the subject of the main verb in other language. For example: "*Deepa pleases Nitu.*" will be translated into Hindi as " n*itu deepa ko pasand kartii hei*" ("*Nitu likes Deepa*."). The verbal object in English "*Nitu*" becomes the subject of the main verb in Hindi

3.2. Promotional deviation

T he modifier is realized as an adverbial phrase in one language but as the main verb in other language. For example: "*Fan is* on" in English, will be translated as "*pankhaa chal rahaa hai*" This means that English modifier "*on*"(an adverb) is realized as the main verb in Hindi.

3.3. Structural deviation

The verbal object is realized as a noun phrase in one language and as a prepositional phrase in other language. For example, the English sentence "*Ram attended the meeting*" will be translated as " *puran sabha mai upashtit tha*". In English " the meeting" is the noun phrase but in Hindi it becomes prepositional phrase *" shaba mein*" (*in the meeting*).

3.4. Conflational deviation

The sense conveyed by a single word in one language requires at least two words of the other language. For example, "*He stabbed me*" will be translated as "u*sne mujhe chaaku se maaraa*". The English word "*stab*" has no one-word equivalent in Hindi, and therefore the introduction of the word "chaaku" was necessitated. Similarly for "*love*", " *swear*"etc.

3.5. Categorial deviation

Changes in category. For example, the predicate is adjectival in one language but nominal in other language. The English sentence "*I am feeling hungry*." will be translated into

Hindi as " *mujhe bhukh lag rahii hai*." In English "*hungry*" is adjective and but in Hindi "*bhukh*" (*hunger*) becomes the noun.

3.6. Lexical deviation

The event is lexically realized as the main verb in one language but as a different verb in other language. Consider the sentence "*They run into the room.*" Its Hindi translation is "*woye daurte huye kamre mein ghus gaye*" The event is lexically realized as the main verb "*run*" in English but as a different verb "*ghus"* (literally (*to enter*)) in Hindi, and "*run*" is used as participle.

3.7. Demotional deviation

A main verb in one language is realized as an adverbial modifier in the other. As shown in [Dorr, 1994] the example "*I like eating.*" will be translated into German "*Ich esse gern* (literal meaning *I eat likingly*)", the word "*like*" is realized as a main verb in English but as an adverbial modifier in German "*gern*". But we have not come across any deviation of this type between English and Hindi.

Some other deviation examples between English and Hindi that we have identified are:

3.7.1. The verb in English language is realized as subject in Hindi. For example, *"It is raining*." in English will be translated as "*baarish ho rahii hai*." In English "rain" is the verb but in Hindi it becomes as a subject "*baarish*".

3.7.2. The prepositional phrase in English becomes the main verb in Hindi. Translation of the sentence *"The train is in motion*." In Hindi is "*gadhi chal rahii hai*". In English "*in motion*" is the prepositional phrase, but in Hindi it becomes the main verb "*chalna*".

3.7.3. The noun in English is realized as the main verb in Hindi. Consider, for example, *"You should give it a try*." Its Hindi translation is " *tumhei kaushish karnii chahiye*". In English "*try*" is noun but in Hindi it becomes the main verb " *kaushish karna*".

3.7.4. The adjective in English changes to the subject in Hindi. For example: " *I am sleepy*." in English will be translated into Hindi as "*mujhe niid aa rahii hai".* The English adjective "*sleepy*" becomes the subject " *niid* " in Hindi

4. Identification of Deviations

The Fundamentals identification of deviations can be achieved through some systematic representations of sentences. A sentence may be represented from two perspectives: syntactic structure and lexical-semantic.

If there is any difference in the role

4.1. Syntactic Structure

Here the constituent words are categorized from their significance in the overall syntax of the given sentence. Some of these categories are: Complementizer (C). This category corresponds to relative pronouns such as *that* in "*the man that I* *saw.*” Inflection (I). This category corresponds to modals such as the word *would* in the sentence “*I would eat cake”*. The term “Inflection” refers to verbal inflection, not other types of inflection.

The other categories are: Verb (V), Noun (N), Adjective (A), and Preposition (P), Adverb (ADV), with their usual significance.

For each category one uses a suffix “P” to denote a phrase and not a word.

Consider for example, the “*Sita went hurriedly to hospital”*. Its structural representation in English (E) and
Hindi (H) are:

E: [$_{CP}$ [$_{IP}$ [$_{NP}$ Sita] [$_{VP}$ [$_{V}$ went] [$_{ADV}$ hurriedly] [$_{PP}$ to [$_{NP}$ hospital]]]]]

H:[$_{CP}$ [$_{IP}$ [$_{NP}$ sita] [$_{VP}$ [$_{ADV}$ jaldi se] [$_{PP}$ [$_{NP}$ asptaal] [$_{P}$ ko [$_{V}$ gayai]]]]]

Where [$_{V}$ went] is syntactic head (verb), [$_{NP}$ Sita] is syntactic subject, [$_{PP}$ to…] is syntactic object and [$_{ADV}$ hurriedly] is the syntactic modifiers. Similarly in Hindi also. Note that the object constituent is itself a syntactic phrase that contains an object, [$_{NP}$ hospital], i.e., structural is recursively defined.

4.2. Lexical-semantic

Here constituent words are analyzed to provide an intermediate representation to the system in a form called Lexical Conceptual Structure (*LCS*). The *LCS* may be achieved by unifying the Root Lexical Conceptual Structures (*RLCS*) of the constituent words.

Lexical Conceptual Structure (*LCS*) is compositional in nature and provides an abstraction for a sentence that is independent of the underlying language. The *LCS* representation of “*Sita went hurriedly to hospital”* is:

[$_{Event}$ GO$_{Loc}$ ([$_{Thing}$ SITA], [$_{Path}$ TO$_{LOC}$ ([$_{Position}$ AT$_{LOC}$ ([$_{Thing}$ SITA],[$_{Location}$ HOSPITAL])])] [$_{manner}$ HURRIEDLY])]

Where GO$_{Loc,}$ is *LCS* head (verb), SITA is *LCS* subject, TO$_{LOC}$ is *LCS* object, and HURRIEDLY is the *LCS* modifier.

Root lexical conceptual structure (*RLCS*) is an un-instantiated *LCS* that is associated with a word definition in the lexicon. For example, the *RLCS* associated with the word “*go*” from above example is:

[$_{Event}$ GO$_{Loc}$ ([$_{Thing}$ X], [$_{Path}$ TO$_{Loc}$ ([$_{Position}$ AT$_{Loc}$ ([$_{Thing}$ X],[$_{Thing}$ Z])])])]

The *LCS* obtained at the end through a series of unification is the final language-independent form of the lexical-semantic representation of a sentence, and serves as the link between source and target languages. Unification of the *RLCS* for “go” (given above) with the *RLCS*’s for “Sita” ([$_{Thing}$ SITA]), “hospital” ([$_{Location}$ HOSPITAL]), and “hurriedly” ([$_{Manner}$ HURRIEDLY]), we get the

Step 2 : relate the syntactic object to the *LCS* subject	O S
Step 3 : relate the syntactic subject to the *LCS* object	S O

final composed *LCS* (*CLCS*) for the sentence “Sita went hurriedly to hospital” as shown earlier.

Once the two representations are obtained, the question of connecting the two different representations arises. A Generalized Linking Routine (*GLR*) is used for this purpose.

Generalized Linking Routine (GLR) correlates the constituents of the syntactic representations to those of the *LCS* representation by the following mappings:

i. Relate the syntactic verb (V) to the *LCS* verb (V). V V
ii. Relate the syntactic subject (S) to the *LCS* subject (S). S S
iii. Relate the syntactic objects (O $_{1}$….,O $_{n}$) to the *LCS* objects (O$_{1}$….,O$_{n}$) O $_{1}$….,O $_{n}$ O$_{1}$….,O$_{n}$
iv. Relate the syntactic adjuncts (M $_{1}$..,M $_{m}$) to the *LCS* modifiers (M$_{1}$...,M$_{m}$) M $_{1}$…,M $_{m}$ M$_{1}$…,M$_{m}$

For example, from our above example we get the following correspondences:

a) V = GO$_{Loc}$ V=[$_{V}$ went];
b) S = Sita S= [$_{NP}$ Sita];
c) O = TO$_{loc}$ O =[$_{pp}$ to …];
d) M = HURRIEDLY M=[$_{ADV}$ hurriedly].

Finally, the lexical-semantic items are systematically related to their respective syntactic categories using *Canonical Syntactic Realization* (*CSR*): For example, an EVENT is a verb (V), a THING is a noun (N), a POPERTY is an adjective (A), a PATH is a preposition (P), TIME, MANNER are adverbs (ADV). Many example sentences are needed to be analyzed to obtain an exhaustive list of *CSR*. We are currently working towards this goal.

The solution of the deviation problem now depends on the *GLR* and the *CSR* information of a sentence. In general, translation deviation occurs when there is an exception either to the *GLR* or to the *CSR* (or to both) in one language but not in the other. This premise allows one to formally define a classification of all possible lexical-semantic

“*Sita went hurriedly to hospital”* is:

deviation that could arise during translation.

5. Identification of Different Diversions

In this section we present the technique for identifying all the six types of deviation that we could find for English and Hindi translation

5.1 Identification of Thematic Deviation

The thematic deviation arises in cases where the *GLR* invokes the following steps of relation in place of steps 2. and step 3. of *GLR*:

An example of thematic deviation (see the sentence "Deepa pleases Nitu") is given in Section 3. The syntactic structure and corresponding CLCS are shown here:

[$_{CP}$ [$_{IP}$ [$_{NP}$ Deepa] [$_{VP}$ [$_{V}$ pleases] [$_{NP}$ Nitu]]]]
[$_{State}$ BEI$_{Ident}$ ([$_{Thing}$ DEEPA],[$_{Position}$ AT$_{Ident}$ ([$_{Thing}$ DEEPA],[$_{Thing}$ NITU])] , [$_{manner}$ LIKINGLY])] [$_{CP}$ [$_{IP}$ [$_{NP}$ nitu] [$_{VP}$ [$_{VP}$ [$_{NP}$ [$_{N}$ deepa][p ko]][$_{V}$ pasand kartii]] [$_{V}$ hai]]]

Here, the object "Nitu" has reversed places with the subject "Deepa" in the Hindi translation. The result is that the object "Nitu" turn into the subject, and the subject "Deepa" turns into the object.

5.2. Identification and Promotional Derivation

Promotional deviation is characterized by promotion of a *LCS* modifier. In such situations, the *LCS* modifier is associated with the structural verb, and the *LCS* verb is then associated with a structural object. Thus, the promotional deviation overrides the *GLR*, invoking the following sets of relations in place of steps 1 and 4 of *GLR:*

Step 1 : relate the *LCS* verb to the syntactic object V O

Step 4 : relate the *LCS* modifier to the syntactic verb M V

The first relation does not mean that V replaces O (if there is a O), but that V retains the same structural relation with O (i.e., O remains an object of V).

An example of structural deviation is given in Section 3. The syntactic structure and corresponding *CLCS* are shown here:

[$_{CP}$ [$_{IP}$ [$_{NP}$ Fan] [$_{VP}$ [$_{V}$ is] [$_{AP}$ on]]]]

[$_{State}$ BEI$_{Ident}$ ([$_{Thing}$ FAN], [$_{Position}$ AT$_{Ident}$ ([$_{Thing}$ FAN],[$_{manner}$ ON)]

[$_{CP}$ [$_{IP}$ [$_{NP}$ pankhaa] [$_{VP}$ [$_{V}$ chal] [$_{V}$ *rahaa hei*]]]]

Here English modifier "*on*"(an adverbial phrase) is realized as the main verb in Hindi.

5.3. Identification of Structural deviation:

Structural deviation differs from the last two deviation types in that it does not alter the positions used in the GLR mapping, but it changes the nature of the relation between the different positions

An example of structural deviation is the case given in section 3. The syntactic structure and corresponding CLCS are shown here:

[CP [IP [NP Ram] [VP [V attended] [NP the meeting]]]]

[$_{Event}$ GO$_{Loc}$ ([$_{Thing}$ RAM], [$_{Path}$ TO$_{Loc}$ ([$_{Position}$ IN$_{Loc}$ ([$_{Thing}$ RAM],[$_{Location}$ MEETING])])])]

[$_{CP}$ [$_{IP}$ [$_{NP}$ ram] [$_{VP}$ [$_{NP}$ [$_{N}$ sabha] [$_{P}$ mein]] [$_{V}$ upasthit tha]]]]

Here, the verbal object is realized as a noun phrase (*the house*) in English and as a prepositional phrase (*sabha mein*) in Hindi.

5.4. Identification of Conflational Deviation

Conflational deviation is another type in which the correspondence is changed. In particular, conflational deviation is characterized by the suppression of a *CLCS* constituent (or the inverse of this process). The constituent generally occurs in LCS object or LCS modifier; thus, the correspondence of either step 3 or step 4 of the *GLR* is changed, depending on which position is conflated. The *LCS* object in the *LCS* does not have a corresponding realization in the syntax.

For the conflational deviation example, given in section 3 ("He stabbed me"), the syntactic structure and corresponding *CLCS* are shown here: [$_{CP}$ [$_{IP}$ [$_{NP}$ He] [$_{VP}$ [$_{V}$ stabbed] [$_{NP}$ me]]]]

[$_{Event}$ CAUSE ([$_{Thing}$ HE], [$_{Event}$ GO $_{Poss}$ ([$_{Thing}$ KNIFE-WOUND],
[$_{Path}$ TOWARD$_{Poss}$ ([$_{Position}$ AT$_{Poss}$ ([$_{Thing}$ KNIFE-WOUND],[$_{Thing}$ ME])])])])]

[$_{CP}$ [$_{IP}$ [$_{NP}$ usne] [$_{VP}$ [$_{NP}$ mujhe] [$_{NP}$ [$_{N}$ chaaku] [$_{p}$ se]][$_{V}$ maaraa]]]]

Here, English uses the single word "*stab*" for the two Hindi words "*chaaku se*" (*knife-wound) and* "*maaraa*" (*kill).*

5.5. Identification of Catogrial Identification:

Lexical deviation is viewed as a side effect of other deviations. Thus, the formulation thereof is considered to be some combination of those given above. For example, in the lexical deviation example mentioned in section 3.,a conflational deviation forces the occurrence of a lexical deviation.

The syntactic structure and corresponding CLCS for this example are shown here:

[$_{CP}$ [$_{IP}$ [$_{NP}$ They] [$_{VP}$ [$_{V}$ run] [$_{PP}$ into [$_{NP}$ the room]]]]]

[$_{Event}$ CAUSE ([$_{Thing}$ THEY], [$_{Event}$ GO $_{Loc}$ ([$_{Thing}$ THEY], [$_{Path}$ TO$_{Loc}$ ([$_{Position}$ IN$_{Loc}$ ([$_{Thing}$ THEY

],[$_{Location}$ ROOM])])])] [$_{manner}$ RUNNINGLY])]
[$_{CP}$ [$_{IP}$ [$_{NP}$ woye] [$_{VP}$ [$_{ADV\ P}$ daurte huye] [$_{PP}$[$_{N}$ kamre] [$_{P}$ mein] [$_{V}$ ghus gaye]]]]
Here the main verb "*run*" in English but as a different verb "*ghus gaye"* (literally (*to enter*)) in Hindi

6. Concluding Remarks

Success of EBMT depends on how efficiently a retrieved translation example can be modified to meet a given translation requirement. Although syntactic rules of the source and target languages are generally helpful, they are not capable of handling exceptional cases called deviations. Hence identification of deviation is an essential component of adaptation in the context of EBMT.
In this work we have provided a systematic scheme for identification of deviations between English and Hindi translations. The scheme depends upon systematic representation of sentences form structural and lexical-semantic points of view. We are still working on a collection of translation examples in order to identify all the possible types of deviations between English and Hindi translations, and how can they be identified through structural and lexical-semantic representation.

Resolution of deviations almost invariably needs special treatment. It can be done by framing appropriate transfer rules, or by using some parameterised mappings that can be applied uniformly across all languages. The parameters can be used to invoke exceptions to GLR and CSR functions in the context of translation deviation. We are presently working on this aspect.

References:

[Nirenburg, 1994] Nirenburg, S., Beale, S., and Domashnev, C., (1994) "A Full-Text Experiment in Example-Based Machine Translation", in: *Proceedings of the International Conference on New Methods in Language Processing, NeMLap, Manchester, UK, 1994, pp: 78-87*.
[Dorr, 1994] Bonnie J. Dorr. (1994), "Machine Translation Deviations: A Formal Description and Proposed Solution", *ACL Vol. 20, No. 4, pp. 597-631*.
[Rao, 1998] Rao, D., Bhattacharya, P., Mamidi, R. (1998)."Natural Language Generation for English to Hindi Human-Aided Machine Translation" *KBCS_1998, Bombay pp. 179-189*
[Dorr, 1993] Bonnie Jean Dorr, (1993). "*Machine Translation: A View from the Lexico*n", The MIT press, USA.