LANGUAGE IN INDIA

Strength for Today and Bright Hope for Tomorrow

Volume 9 : 10 October 2009

ISSN 1930-2940

Will Sentences Have Divergence Upon Translation?
A Corpus-Evidence Based Solution for Example Based Approach

Deepa Gupta, B.A (Hons), M.Sc., (Maths), Ph.D.

Will Sentences Have Divergence Upon Translation? : A Corpus-Evidence Based Solution for Example Based Approach

Deepa Gupta, B.A (Hons), M.Sc(Maths), Ph.D
Amrita Vishwa Vidyapeetham University
Department of Mathematics
Amrita School of Engineering
Kasavanahalli, Bangalore -560 035
Karnataka, India
deepag_iitd@yahoo.com

Abstract

This paper presents a corpus-evidence based scheme for deciding whether the translation of an English sentence into Hindi will involve *divergence*. Divergence is the phenomenon when sentences of similar structure in the source language do not translate into structurally similar sentences in the target language. Divergence assumes special significance in the domain of Example Based Machine Translation (EBMT) where translation of a given sentence is generated by first *retrieving* translation example(s) of similar sentence(s) from the system's example base, and then by *adapting* them suitably to meet the requirements of the present input sentence. Surely, occurrence of divergence poses a great hindrance in efficient adaptation

of retrieved sentences. A possible remedy may lie in dividing the example base of an EBMT system into two parts: examples of *normal* translation, in one, and examples involving *divergence* in the other, so that given an input, the retrieval can be made from the appropriate part of the example base. But success of this scheme depends heavily on the system's ability to judge a priori whether translation of a given input will involve divergence. The task, however, is not straightforward as occurrence of divergence does not follow any rules that make their prior identification simple. The technique proposed here is aimed at achieving this goal. The scheme is explained and illustrated in the context of English to Hindi EBMT.

1 Introduction

Dealing with divergence is one major difficulty of any translation system. Typically, in a translation the structure of the translated sentence is guided by the syntactic and semantic properties of the target language. If upon translation the Parts of Speech (POS) and Functional Tags (FT) of the constituent words of the source language sentence do not undergo any changes then we term it as a *normal* translation. However, there are occasions when the structure of the translated sentence deviates from this normal structure. Such exceptions are called translation *divergences* [4]. Consider, for example, the English sentences "It is running" and "It is raining". Although these two sentences are structurally very similar, their Hindi translations are structurally very different. The first sentence is translated as "*wah* (it) *bhaag* (run) *rahaa* (..ing) *hai* (is)", which is a normal translation. But the second one is translated as "*baarish* (rain) *ho* (be) *rahii* (..ing) *hai* (is)". The second example is a clear case of divergence, where the subject of the Hindi sentence is realized from the verb of the English sentence.

Translation divergence has heavy bearings on Example Based Machine Translation (EBMT). In an EBMT system the translation for a given input sentence is generated by retrieving the translation of a similar sentence from the system example base, and then modifying (adapting) them to suit the requirements of the current input sentence [8] [1]. Selection of the right past example is, therefore, extremely important for successful EBMT. The need arises primarily in the following two scenarios:

- The past example that is retrieved for carrying out the task of adaptation has a normal translation, but translation of the input sentence should involve divergence.
- The translation of the retrieved example involves divergence, whereas the input sentence should have a normal translation.

In both the situations the retrieved example may not be helpful in generating the translation of the given input, and consequently, developing efficient adaptation scheme becomes extremely difficult.

A possible solution may lie in separating the example base (EB) into two parts: Divergence EB and Normal EB so that given an input sentence retrieval can be made from the appropriate part of the example base. However, this scheme can work successfully only if the EBMT system has the capability to judge from the input sentence itself whether its translation will involve any divergence. But making such a decision is not straightforward since occurrence of divergence does not follow any patterns or rules. In fact, a divergence may be induced by various factors, such as, structure of the input sentence, semantics of its constituent words etc. In this work we propose a *corpus-evidence* based approach to deal with this difficulty. Under this scheme, upon receiving an input sentence, a system looks into its example base to glean evidences in support/against any possible type of divergence. Based on these evidences the

system decides whether the retrieval has to be made from the normal EB, or from the divergence EB.

A critical look at machine translation suggests that EBMT has been studied extensively as a major paradigm for machine translation over the last decade and more [2]. At the same time literature is replete with works on translation divergence, and its identification, resolution etc. However, the works on these two aspects of machine translation have progressed somewhat independently. No significant work has so far been found regarding how divergence can be dealt with efficiently in an EBMT framework. The proposed work aims at bridging this gap. Since divergence is a language-dependent phenomenon, we have concentrated on a specific source and target language pair, English and Hindi, for this work.

Divergence in English to Hindi translation has been studied thoroughly in some of our earlier works ([5], [6], [7]). With respect to English to Hindi translation, seven different types of divergence have been identified. These are *structural*, *categorial*, *conflational*, *demotional*, *pronominal*, *nominal* and *possessional*. Of the seven types, possessional divergence is somewhat different in nature as unlike the other six, its occurrence depends upon more than one Functional Tag of the sentence. The scheme in its present form cannot handle possessional divergence efficiently. Hence we exclude possessional divergence from the present discussion. The algorithm proposed here, therefore, works with respect to the first six types of divergence. For convenience of presentation we denote them as d_1, d_2, d_3, d_4, d_5 and d_6, respectively.

Barring structural divergence (d_1) all of the other five types of divergence (i.e. d_2,...,d_6) have further been classified into several sub-types depending upon the variations in the role of different functional tags upon translation to Hindi. Appendix-A gives a brief description of all the six divergence types mentioned

above, and their sub-types. It further provides the necessary FT-features that the source langauge (English) sentences should have in order that a particular type/sub-type of divergence may occur. This, however, does not mean that any sentence having those FT-features will necessarily produce a divergence upon translation. As a consequence, mere examination of the FTs of an input sentence cannot ascertain whether its translation will induce any divergence or not. Hence more evidences need to be considered. In this work we describe all these evidences and how they are to be used for making a priori decision regarding whether the input English sentence will involve any divergence upon translation to Hindi.

This paper is organised in the following way. Section 2 explains the different types of corpus-based evidences that are used by the proposed approach. Most of these evidences are formulated by analysing a parallel corpus comprising more than 4000 sentences collected from various sources, such as, children's stories, translation books, advertisement materials and official letters. Sections 3 explain how different evidences are generated and combined to arrive at a final decision regarding an input. Section 4 provides illustrations of the scheme, and experimental results.

2 Corpus-Based Evidences and Their Use in Divergence Identification

The proposed scheme make use of three different types of evidence to decide whether a given input sentence, will have a normal translation, or whether it will involve one (or more) type of divergence when translated into Hindi. These evidences are used in succession to obtain the overall evidence in support of divergence(s)/non-divergence in the translation of the input sentence. These

three steps are explained below:

Step1: Here Functional Tags (FTs) of the constituent words of the input sentence are used to determine the divergence types that cannot certainly occur in the translation of that sentence. The output of this step is a set D of divergence types that may possibly occur in the translation of a given input sentence.

Step2: Here semantic similarities of constituent word(s) of input sentence with constituent words of sentences in the divergence EB and the normal EB are determined. Depending on the occurrence of similar words in the divergence and/or normal EB the scheme decides whether upon translation the input sentence may induce any divergence.

Step3: Some times the above two steps may suggest more than one type of divergence. In such a situation the algorithm should consult its knowledge base to ascertain which combinations of divergence type are possible in the translation of a single sentence.

A scrutiny of our example base, and examination of the syntactic rules of the Hindi grammar suggest that only the following combinations of divergence are possible with respect to English to Hindi translation:

1. structural (d_1) and conflational (d_3)
2. conflational (d_3) and demotional (d_4)
3. categorical (d_2) and pronominal (d_5).

This knowledge is stored in a set CD := $\{\{d_1, d_3\}, \{d_3, d_4\}, \{d_2, d_5\}\}$. The possible combinations of divergence can be used as evidence to rule out any suggestions given by the earlier two steps that do not conform with the knowledge stored in the set CD described just above.

f_1 :	Root form of the main verb is "be"
f_2 :	To-infinitive form of a verb is present
f_3 :	Root form of the main verb is not "be/have"
f_4 :	Subject is present
f_5:	Object is present
f_6 :	Subjective complement (SC) is present
f_7 :	Subjective complement is adjective
f_8 :	Subject of the sentence is "it"
f_9 :	Verb complement(VC) is present and is a PP
f_{10}:	Predicative adjunct (PA) is present

Table 1: FT-features Instrumental for Creating Divergence

The following subsections elaborate the above steps.

2.1 Roles of Different Functional Tags

Analysis of the divergence examples suggests that for each divergence type to occur the underlying sentence needs to have some specific functional tags (FT) and/or some specific attributes of these FTs. We call them together *FT-features* of a sentence. Appendix-A contains this information for each divergence type and sub-type. Considering all the divergences together we found that ten different FT-features are, in particular, useful for identification of divergence. Table 1 provides a list of these features, which we label as f_1, f_2, ..., f_{10}.
With respect to a particular type of divergence, an FT-feature may have one of the following three roles:

- Its presence in the input sentence is necessary should the corresponding divergence occur.
- It should necessarily be absent in the input sentence if the corresponding divergence is to occur.
- The FT-feature has no role in the occurrence of the corresponding divergence.

We denote the above three possibilities as P (present), A (absent), and X (don't care). Table 2 gives the roles of the 10 FT-features discussed above in the occurrence of the different types of divergence and their sub-types. We call the table as "Relevance Table".

	sub-type	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
d_1	-	X	X	P	X	P	A	A	X	A	A
d_2	sub-type1	P	X	A	P	A	P	X	X	A	A
	sub-tyep2	P	X	A	P	A	P	X	X	A	A
	sub-tyep3	P	X	A	P	A	A	X	X	A	P
	sub-tyep4	P	X	A	P	A	A	X	X	A	P
d_3	sub-type1	A	X	P	X	X	X	X	X	X	A
	sub-type2	A	X	P	P	X	X	X	X	X	A
d_4	sub-type1	A	X	P	P	P	A	A	X	A	A
	sub-type2	A	X	P	P	A	A	A	X	P	A
	sub-type3	A	X	P	P	P	A	A	X	A	A
	sub-type4	A	X	P	P	P	A	A	X	A	A
d_5	sub-type1	P	X	A	P	A	P	X	P	X	A
	sub-type2	A	X	P	P	A	X	X	P	X	A
	sub-type3	P	P	A	P	A	P	X	P	A	A
d_6	sub-type1	P	X	A	P	A	P	P	X	A	A
	sub-type2	A	X	P	P	A	P	P	X	A	A

Table 2: Relevance of FTs in different divergence type

Each row of the Relevance Table provides the necessary conditions on the FT-features of an input sentence in order that the corresponding divergence may occur. The advantage of this evidence is that it helps in quick discarding of those types of divergence that *cannot* occur in the translation of the given input sentence.

The information given in Table 2 may be used in the following way. Given an input sentence, the algorithm first extracts the values for the 10 FT-features, f_j, j = 1, 2, ..., 10, from the sentence. These values are then compared with the row entries of the Relevance Table. If the FT-features of the sentence conform with the entries of some particular row, then evidence is obtained towards occurrence of that particular divergence for which this row corresponds to one of the sub-

types. If a particular sentence has evidence supporting more than one divergence then all these possible divergence types are to be considered for step 2 of the algorithm. This set of possible divergence types for a given input is denoted as D.

For illustration, consider the following input sentence: Ram is friendly to me. As the sentence is parsed (with some unnecessary components edited) one may get the following:
@SUBJ <Proper> N NOM SG "Ram", @+FMAINV V PRES "be", @PCOMPL-S A ABS "friendly" , @ADVL PREP "to", @<P PRON PERS SG1 "i" "< $. >"

The notations used here are from ENGCG parser(*http://www. lingsoft. fi /cgi-bin/engcg*). Appendix-B provides a short description of the functional tags used in the parsed output. We can summarize the parsed version as follows. Of the ten FT-features discussed above (see Table 1) only four are present in the above sentence. These are:

- f_1 – because the main verb of the sentence is "be".
- f_4 – since the sentence has a subject, viz. "Ram".
- f_6 – as an SC "friendly" is present in the sentence.
- f_7 – since the SC of this sentence is an adjective.

Thus in the Hindi translation of this sentence only those divergence sub-types can occur for which the entries corresponding to FT-features f_1, f_4, f_6, and f_7 are either "P" or "X". For the other FT-features the entries have to be either "A" or "X". This algorithm assumes that occurrence of a particular divergence type is possible only if at least one of its sub-types satisfies the above conditions. Thus for the above input sentence the possible divergences are:

- Categorial (d_2), since sub-types 1 and 2 conform with the above requirements.

- Nominal(d_6), since sub-type 1 satisfies the above requirements.

Also note that sub-type 1 of d_5 has values either "P" or "X" for the FT-features f_1, f_4, f_6, and f_7. But divergence d_5 cannot occur in this case as the sub-type has an extra requirements that FT-feature f_8 should also be present, which is not true for this sentence. Therefore, the output of this step is the set $D = \{d_2, d_6\}$.

It however should be noted that the FT-features specified in the Relevance Table do not provide conclusive evidence towards the presence of some particular divergence type. For example, consider the following two sentences.

Example (A):

She is in trouble. $\sim$ *wah* (she) *musiibat* (trouble) *mein* (in) *hai* (is)

She is in tears. $\sim$ *wah* (she) *ro* (cry) *rahii* (..ing) *hai* (is)

Since both the sentences given in Example (A) have the same FT-features, i.e., f_1, f_4 and f_{10}, the Relevance Table gives evidence supporting categorial divergence d_2 (check the rows for sub-types 3 and 4) for both the sentences. But of the two sentences the translation of the first one is a normal one. It is only the second sentence that involves categorial divergence upon translation to Hindi. Thus, to determine the possible divergence type(s) in a sentence, only the FT-features cannot be taken as the sole evidence, and more evidences need to be sought.

From the above example, it can be surmised that it is the prepositional phrase "in tears" that is instrumental for causing the categorial divergence in the second sentence. In general, corresponding to each divergence type one can associate some functional tags that are instrumental for causing the divergence. We call it the *Problematic FT* of the corresponding divergence type. Table

3 provides the Problematic FT corresponding to all the six divergence types relevant in the context of English to Hindi translation. This table has been obtained by examining the sentences in our example base.

Table 3 is to be used in the following way. If the FT-features of a given input conform with the requirements of a particular divergence type (as given in the Relevance Table) then the corresponding problematic FT in the sentence needs to be examined more carefully. Since both the sentences of Example (A) have the structures required for categorial divergence, Table 3 suggests that to gather more evidence the scheme should concentrate on the SC or PA of the sentences.

Divergence Type	**Problematic FT**
Structural	Main Verb
Categorial	Subjective Complement (SC: adjective, noun) or Predicative Adjunct (adverb, PP)
Conflational	Main Verb
Demotional	Main Verb
Pronominal	Main Verb or Subjective Complement (adjective, noun)
Nominal	SC (adjective)

Table 3: FT of problematic words for each divergence type

In this respect one major difficulty is that a particular word may convey different senses in different context even if it is under the same FT. For example, consider the two sentences and their Hindi translations given in Example (B) below:

Example (B):

Mohan beat the drum in the school. ∼

Mohan ne	*vidyaalay*	*mein*	*drum*	*bajaayaa*
(Mohan)	(school)	(in)	(drum)	(beat)

Agassi beat Becker in the final. ∼

Agassi ne	*final*	*mein*	*Becker ko*	*haraayaa*
(Agassi)	(final)	(in)	(Becker)	(beat)

Here, the first one is an example of normal translation, while the second one is a case of structural divergence because of the introduction of the preposition "*ko*" in the object of the Hindi sentence. A careful examination suggests that although the main verb of both the sentences is "beat", its translation causes divergence when used in a particular sense, but not when used in some other sense. By referring to WordNet 2.0 {*http://www.cogsci.princeton.edu/cgi-bin/webwn*} one may find that the first sentence has the 6^{th} sense of the word "beat", which is "*to make a rhythmic sound*"; while the second sentence has the 1^{st} sense of the word "beat", which is "*to come out better in a competition, race, or conflict*". Therefore, while dealing with words one needs to pay attention to the particular sense in which a word is being used – in some senses it may cause divergence, and in some other senses it may not induce any divergence at all.

Since an exhaustive list of words (along with their relevant senses) that lead to divergence is impossible to make, the proposed algorithm tries to gather more evidences by using the semantic similarity of the constituent words to the word senses that are already known to cause divergence, or known to deliver a normal translation.

In order to achieve the above, two dictionaries have been created: *Problematic Sense Dictionary* (PSD) and *Normal Sense Dictionary* (NSD). The PSD contains the words along with their senses that have been found to cause divergence. Similarly, the NSD contains the words along with their senses for which normal translation has been observed.

These dictionaries are further grouped into six sections – a section corresponding to each divergence type. Section PSD_i contains problematic words,

Divergence type	(d_i)	No. of words in PSD$_i$	No. of words in NSD$_i$
Structural	(d_1)	163	1078
Categorial	(d_2)	57	167
Conflational	(d_3)	43	997
Demotional	(d_4)	66	1422
Pronominal	(d_5)	75	170
Nominal	(d_6)	12	97
Total		416	3931

Table 4: Frequency of words in different sections

occurring in sentences whose translations involve divergence of type d_i. Similarly, section NSD$_i$ contains problematic words of sentences having the FT-features as required for divergence type d_i (as specified in the Relevance Table), but actually having a normal translation. Table 4 gives the number of words in each section of the PSD and the NSD that is currently present in our example base.

PSD$_1$	**NSD$_1$**	**PSD$_2$**	**NSD$_2$**
Attend#v#1	Beat#v#6	Afraid#a#1	Brave#a#1
Beat#v#1	Do#v#13	Friendly#a#4	Good#a#1
Love#v#3	Eat #v#4	On#r#2	Illusion#n#2
Marry#v#1	Purchase#v#1	Pain#n#1	Monitor#n#2
Occupy#v#4	See#v#1	Tear#n#1	Trouble#n#1
...	...	...	...
PSD$_3$	**NSD$_3$**	**PSD$_4$**	**NSD$_4$**
Face#v#3	Agree#v#4	Belong#v#1	Continue#v#9
Look#v#5	Feel#v#4	Face#v#3	Ride#v#9
Resemble#v#1	Go#v#10	Front#v#1	Sell#v#2
Rush#v#4	Look#v#3	Smell#v#2	Solve#v#1
Stab#v#1	Solve#v#1	Suffice#v#1	Walk#v#6
...	...	...	...
PSD$_5$	**NSD$_5$**	**PSD$_6$**	**NSD$_6$**
Freeze#v#6	Bright#a#10	Cold#a#1	Dull#a#4
Humid#a#1	Light#a#1	Hot#a#1	Good#a#1
Morning#n#3	Plain#a#2	Hungry#a#1	Happy#a#2
Rain#v#1	Shiny#a#3	Sleepy#a#1	Helpful#a#1
Winter#n#1	Wrong#a#1	Thirsty#a#2	Innocent#a#4
...	...	...	...

Table 5: PSD/NSD Schematic Representations

Each PSD/NSD entry contains along with the relevant word, its part of speech and appropriate sense number (as given by WordNet 2.0). Table 5 shows some entries corresponding to each PSD_i and NSD_i, i=1,2,...6. The entries are stored in the format *word#pos#k*, where *pos* stands for the particular Part of Speech, which can be one of n, v, a or r (corresponding to noun, verb, adjective and adverb, respectively), and k stands for the sense number.

For illustration, consider the two sentences given in Example (A). Both of them have the structure required for categorial divergence i.e. d_2. Problematic FT for this divergence type is the predicative adjunct (PA), which is a prepositional phrase. Hence, in PSD_2 and NSD_2 we store *tears#n#1* and *trouble#n#1*, respectively. Similarly, corresponding to Example (B) where the relevant divergence is structural i.e. d_1, the entries in PSD_1 and NSD_1 are *beat#v#1* and *beat#v#6*, respectively.

In order to ascertain whether a given input sentence may have a divergence d_i the proposed scheme proceeds as follows. It first identifies the problematic word a_i of the sentence corresponding to the divergence d_i. The evidence is collected on the basis of four parameters, viz. $sim(a_i, w_i)$, $s(d_i)$, $sim(a_i, w_i')$ and $s(n_i)$, as described below:

1. $sim(a_i, w_i)$ gives the maximum similarity score between a_i and the words in PSD_i, where $sim(x, y)$ denote the semantic similarity between two words x and y (see Appendix-C).

2. The quantity $s(d_i)$, corresponding to divergence type d_i is defined as follows:

$$s(d_i) = \begin{cases} 0 & if \;\; x_i = 0 \\ \frac{1}{2}\left(\frac{x_i}{c_i} + \frac{c_i}{S}\right) & otherwise. \end{cases} \quad ...(1)$$

where c_i, x_i and S are as follows:

(a) c_i is the total number of entries in PSD_i (given in Table 4);

(b) x_i is the number of words in PSD_i that are semantically similar to a_i;

(c) S is the total number of words in the PSD. Note that currently the total number of words in PSD is 416 (see Table 4);

3. The quantity $sim(a_i, w_i')$ is similar to $sim(a_i, w_i)$. While computing $sim(a_i, w_i')$, the scheme will use NSD_i and NSD instead of PSD_i and PSD.

4. The quantity $s(n_i)$ is similar to $s(d_i)$, and is calculated using NSD_i and NSD. The value used for S here is the cardinality of the NSD which is 3931 (see Table 4).

These four quantities are used to determine the possibility of occurrence of divergence d_i in the translation of the given input sentence.

3 The Proposed Approach

In order to determine whether a given input sentence, say e, may involve some divergence upon translation, the evidences mentioned in previous section are used in the following way. First the input sentence e is parsed, and then using the Relevance Table a set D is determined that contains the divergence types that may possibly occur in the translation of e. For each possible divergence type $d_i \in D$ the problematic word a_i is extracted from the sentence e. From PSD_i, the word w_i is retrieved that is semantically most similar to a_i. The subsequent steps depend upon the value of $sim(a_i, w_i)$. If the value is 1, that implies that a_i is present in PSD_i. On the other hand, a small value of $sim(a_i, w_i)$ implies that there is not enough evidence in support of divergence d_i. Hence

it may be concluded that divergence d_i will not occur in the translation of e. Note that whether the value of $sim(a_i, w_i)$ is sufficiently small is determined by comparing it with a threshold t, which is to be determined experimentally from the corpus. If the value of $sim(a_i, w_i)$ is between t and 1, then some evidence in support of divergence d_i is obtained. In order to make a conclusion from this point the algorithm now refers to NSD_i to obtain the word w_i' that is semantically most similar to a_i. Depending upon the values of $sim(a_i, w_i)$ and $sim(a_i, w_i')$, a decision is taken regarding whether the translation of e will involve divergence d_i or not. Based on this decision, the retrieval is to be made from the appropriate part of the example base i.e. the Divergence EB or Normal EB.

The overall scheme is explained below which involves four major steps as follows:

Step 1: At this stage, the input sentence e is parsed and its FT-features are obtained. From these FT-features, using Table 2, the set D of possible divergence types is determined.

The main objective now is to determine the divergence types, out of all the $d_i \in D$, which have positive evidence supporting them to happen in the translation of e. Steps 2 and 3 are designed for this purpose. A set of flags, Flag_i, corresponding to each $d_i \in D$ is used to store this information. Initially each of these flags is set to –1. Step 2 and Step 3 are now carried out for each $d_i \in D$ in order to reassign the value of Flag_i. At each iteration the next d_i with the minimum index i is chosen such that Flag_i is -1.

Step 2: From the input sentence e the problematic word a_i corresponding to divergence d_i (see Section 2) is determined. The set W_i comprising of words belonging to PSD_i and having positive semantic similarity score with a_i is determined. Thus $W_i = \{b : b \in \text{PSD}_i \text{ and } sim(\text{a}_i, b) > 0\}$. From W_i the word

w_i is obtained such that $sim(a_i,\ w_i) = \max\ sim(a_i,\ b) \quad \forall\ b \in \mathrm{W}_i$. If W_i is empty then $sim(a_i,\ w_i)$ is considered to be 0. Depending on the similarity score $sim(a_i,\ w_i)$ decision is taken regarding d_i, as follows.

Case 2a: If $sim(a_i, w_i) = 1$, then set $\mathrm{Flag}_i = 1$. This is because the condition implies that the word a_i is present in PSD_i. Hence this sentence will certainly have divergence d_i upon translation. Therefore Flag_i is set to 1.

Case 2b: This case occurs when $a_i \notin$ PSD. But if a_i is a noun or verb, and further a_i is a *coordinate term* of w_i (i.e. according to WordNet terminology, a_i and w_i have the same hypernym), then it can be decided that a_i will not create divergence of type d_i upon translation. This is because all those coordinate terms of w_i that may cause divergence are already stored in the PSD. Therefore Flag_i is set to 0.

Case 2c: If $sim(\mathrm{a}_i, w_i) < t$, where t is some pre-defined threshold, then too it may be decided that a_i will not cause divergence d_i. Consequently, Flag_i is set to 0. The main difficulty here is to decide upon the right value for the threshold t. After a sequence of experiments with different values for t, we found that the best results are obtained for $t = 0.5$. However, since this value is corpus dependent, for other corpora the value of t should be determined experimentally.

Since in all the three cases above the scheme arrives at a decision regarding the divergence type d_i, computation may skip Step 3 and go to Step 4 directly. But there may be cases when the similarity score $sim(a_i,\ w_i)$ lies between t and 1. In these cases, as mentioned above, the NSD has to be referred to. Hence Step 3 is executed.

Step 3: Here, first the set $\mathrm{W}'_i = \{b\ |b \in \mathrm{NSD}_i$ and $sim(a_i, b) > 0)$ is computed. From this set the word w'_i is picked such that $sim(a_i,\ w'_i) = \max\ sim(a_i, b) \quad \forall b \in W'_i$. If W'_i is empty then $sim(a_i,\ w'_i)$ is considered to be 0. Depending on $sim(\mathrm{a}_i, w'_i)$, one of the following cases is executed.

Case 3a*:* If $sim(a_i, w_i') = 0$ then it implies that there is no evidence that the word will lead to normal translation. Consequently, Flag_i is set to 1 indicating that divergence d_i has a positive chance of happening.

Case 3b: If $sim(a_i, w_i') = 1$ then the evidence suggests that the word a_i should provide a normal translation to the sentence, and there is no possibility of divergence d_i to occur in the translation of this sentence. Consequently, Flag_i is set to 0.

Case 3c: Decision making becomes most difficult when $0 < sim(a_i, w_i') < 1$. This implies that words sufficiently similar to a_i exist neither in the PSD nor in the NSD. Thus, any decision about divergence/ non-divergence cannot be taken yet.

In this case the scheme proposes to look into *how many* words similar to a_i, are available in PSD_i and NSD_i. This evidence is given by score $s(d_i)$ and $s(n_i)$ computed using formula (1) (Given in Section 2). Finally, similarity scores $sim(a_i, w_i)$ and $sim(a_i, w_i')$ are combined with $s(d_i)$ and $s(n_i)$ respectively, to take into consideration the importance of both the evidences. If evidence supporting divergence d_i is more then the value of Flag_i is set to 1 otherwise it is set to 0. Thus, in this case, following computations are performed:

- Compute $s(d_i)$ and $s(n_i)$.
- Determine $m(d_i) := 1/2 * (s(d_i) + sim(a_i, w_i))$, and

  $m(n_i) := 1/2 * (s(n_i) + sim(a_i, w_i'))$.
- If $m(d_i) > m(n_i)$ Then

  Set $\text{Flag}_i = 1$; GO TO Step 4.

  Else If $m(d_i) < m(n_i)$

  Set $\text{Flag}_i = 0$; GO TO Step 4.

  Else $\text{Flag}_i = 1/2$; Break;

The last case refers to a rare situation when $m(d_i)$ and $m(n_i)$ are equal. In this case the algorithm cannot recommend whether the translation will involve divergence d_i, or will it be normal. In such a situation the system can at best pick the most similar examples from both normal EB and divergence EB, and leave it to the user to make the final decision. Therefore, in such cases, the Flag_i is set to 1/2.

Once the evidences supporting/against all divergence types $d_i \in D$ are obtained, that is the value of $\text{Flag}_i \; \forall \; d_i \in D$ is determined, Step 4 is performed to make a final decision regarding possible divergence types in the translation of the given input e. Here it should be noted that the value of $\text{Flag}_i = 0$, implies that e cannot have divergence d_i; while value of $\text{Flag}_i = 1$ implies that upon translation e may have divergence d_i. A set D' is constituted, such that $D' = \{d_i \in D \text{ and } \text{Flag}_i = 1\}$, i.e. D' stores all those d_i's for which positive evidences are obtained.

Step 4: The final decision is computed in the following way.

Case 4a: If $D' = \phi$, then the conclusion is that sufficient evidence has not been obtained for any of the divergence types. Hence, the decision is that the translation of the input sentence e will not involve any divergence.

Case 4b: If $|D'| = 1$, i.e. $D' = \{d_k\}$. This implies that evidence is obtained in support of just one divergence type d_k. The algorithm therefore decides that the translation of the input sentence will have divergence d_k.

Case 4c: If $|D'| > 1$, it implies that there is a possibility of more than one type of divergence. The algorithm therefore seeks further evidence to make any decision. The evidence provided by CD (Section 2) may be used here. A set $\text{C} = \{\{d_i, d_j\} \in \text{CD} \mid \; d_i, d_j \in \text{D}'\}$ is constructed. Depending upon the |C|, further decision has to be taken in the following way.

- If $|\text{C}| = 0$, it implies that no permissible combination has been found. In

this case, the algorithm computes $s(d_i)$ and $m(d_i)$ $\forall \quad d_i \in D'$ as is in Case 3c. The algorithm concludes that the translation of the input sentence will have divergences d_k, where k is such that $m(d_k) = \max_{d_i \in D'} \{m(d_i)\}$.

- If $|C| = 1$, it implies that there is evidence for only one permissible combination. Let it be $\{d_k, d_l\}$. The algorithm suggests that the input sentence e will involve both divergence d_k and d_l upon translation to Hindi.

- If $|C| > 1$, that is, if the evidences are obtained in support of more than one permissible combination of divergences, then the scheme needs to select the most likely combination of them. It therefore determines the quantity $1/2 * (m(d_i) + m(d_j))$ $\forall$ combinations $\{d_i, d_j\} \in C$. The scheme recommends that combination of divergences for which this quantity is maximum.

The flowchart of the proposed scheme is given in Figures 1 and 2.

4 Illustrations and Experimental Results

In this section we first illustrate with examples how the above algorithm works towards prior identification of divergence, if any, in translation from English to Hindi. The examples considered are increasingly difficult in nature. Later in subsection 4.4 a consolidated result of several experiments is presented, and certain limitations of the said algorithm are discussed.

4.1 Illustration 1

Consider the input sentence: I am feeling hungry.

The parsed version of the above sentence is: @SUBJ PRON PERS SG1 "i", @+FAUXV V PRES "be", @-FMAINV PCP1 "feel", @PCOMPL-S A ABS "hungry" < $.>.

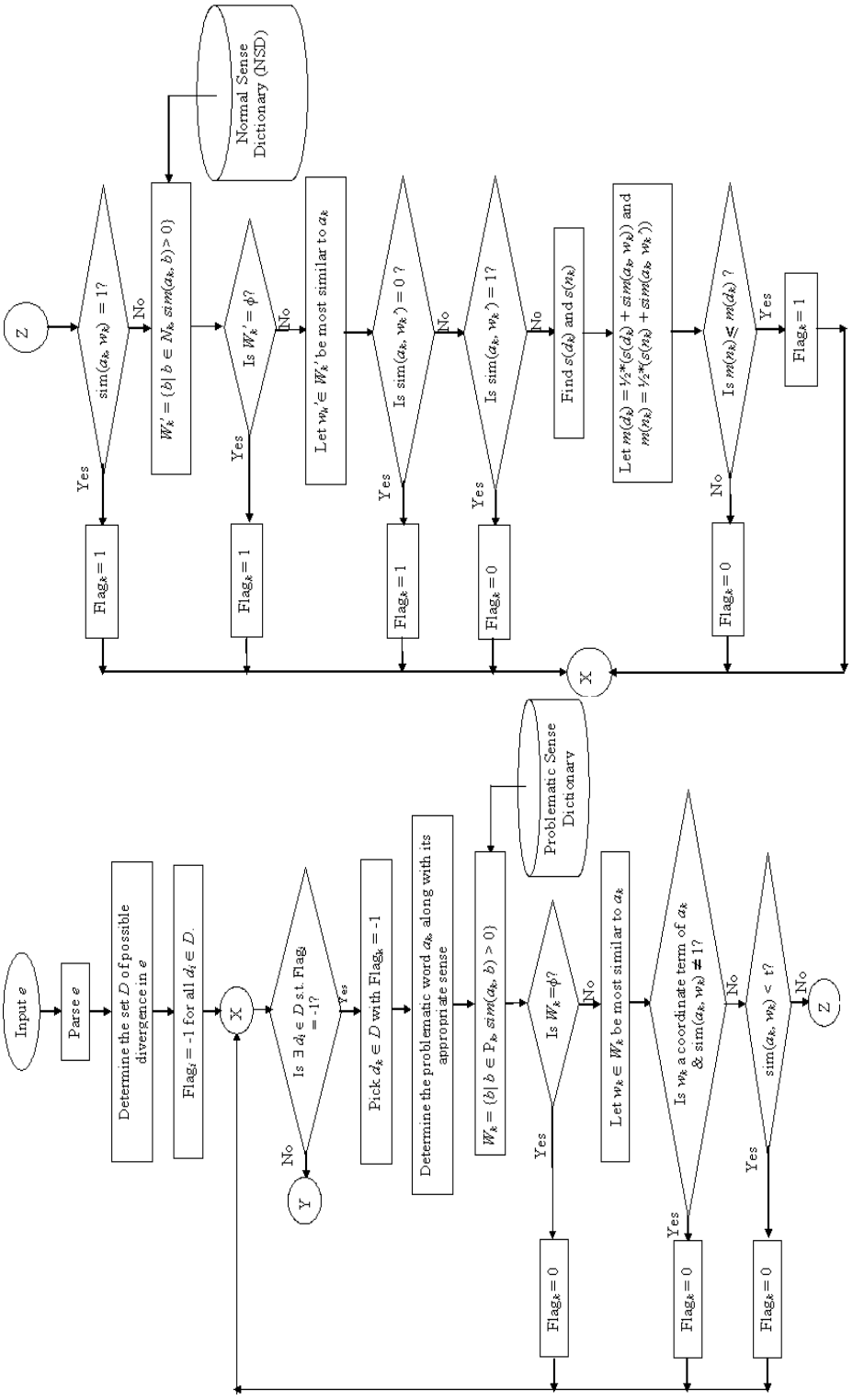

Figure 1: Schematic Diagram of the Proposed Algorithm

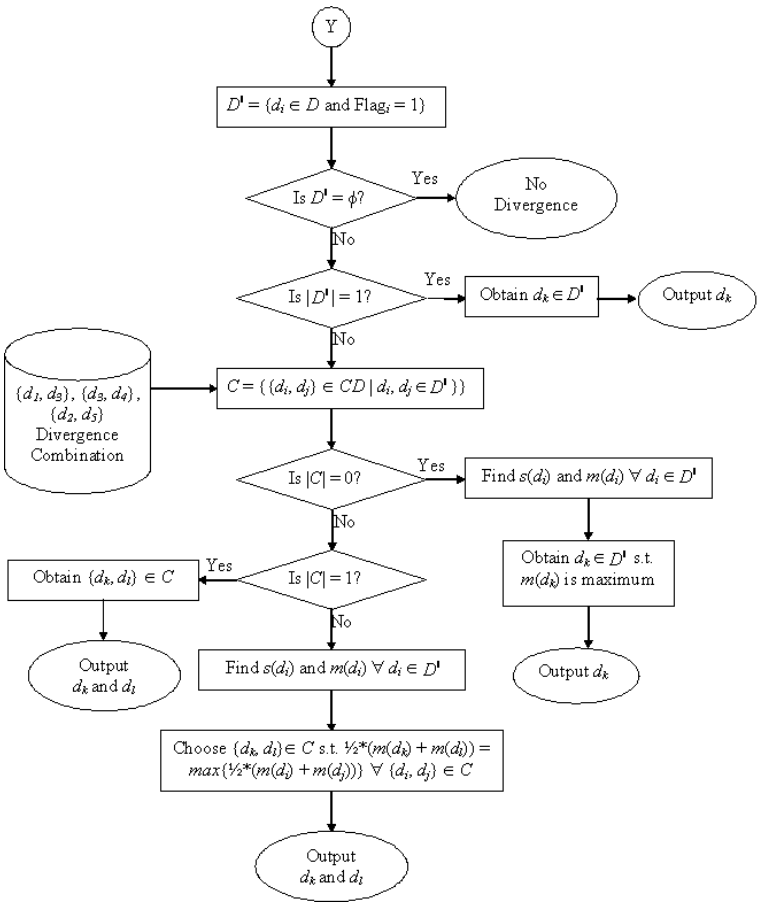

Figure 2: Continuation of the *Figure* 1

Of the ten FT-features (see Table 1) only four are present in the above sentence. These are:

- f_3 – since the main verb (feel) of the sentence is not "be" or "have".
- f_4 – as the sentence has a subject, viz. "I".
- f_6 – because the sentence has an SC.
- f_7 – since the SC of this sentence is an adjective (hungry).

Note that the FT-features of the given input sentence conform with both the sub-types of d_3 and only sub-type 2 of d_6 (see Table 2). Hence the set D of possible divergence types is obtained as D=$\{d_3, d_6\}$ which are conflational and nominal types of divergence, respectively. Therefore, evidences need to be collected for both of the divergence types.

Evidences for conflational divergence (d_3):

Table 3 suggests that the problematic word for d_3 is the main verb i.e. "feel". WordNet 2.0 provides thirteen different senses for the word "feel" when used as a verb, such as:

- *sense1*: feel, experience – undergo an emotional sensation
- *sense2*: find, feel – come to believe on the basis of emotion, intuitions, or indefinite grounds
- *sense3*: feel, sense – perceive by a physical sensation, e.g., coming from the skin or muscles

For the given input sentence the appropriate sense is *sense1*. Thus a_3 is *feel#v#1*. A scrutiny of PSD_3 reveals that it contains no words w such that similarity $sim(w, a_3)$>0. Thus $W_3 = \phi$, and therefore, $Flag_3$ is set to 0.

Evidences for nominal divergence (d_6):

Problematic FT for d_6 is "*Subjectival complement (Adjective)*". Hence the problematic word of the input sentence is "*hungry*". WordNet 2.0 provides two senses for "*hungry*" of which the first one "*feeling hunger*" is appropriate in this case. Thus, the problematic word is a_6 which is *hungry#a#1*. PSD_6 is then scrutinized to find the word semantically most similar to a_6. It is found that PSD_6 already contains *hungry#a#1*. Therefore w_6 is same as a_6 and hence similarity score is 1. Thus $Flag_6$ is set to 1.

Now the set D' is constructed as $D' = \{d_i \in D: \text{Flag}_i = 1\}$. Evidently for the given input sentence D' contains a single element d_6. Thus the algorithm suggests that the above input sentence will cause *nominal divergence* upon translation to Hindi, which is a correct decision.

4.2 Illustration 2

Consider, the input sentence is: She is in a dilemma.
Its parsed version is @SUBJ PRON PERS FEM SG3 "she", @+FMAINV V PRES "be", @ADVL PREP "in", @<P N SG "dilemma" <$.>.

The FT-features present in this sentence are:

- f_1 – as the root form of the main verb is "be".
- f_4 – because the sentence has a subject, viz. "she".
- f_{10} – since the sentence has a PA, viz. "dilemma".

Using the Relevance Table the set D of possible divergence types is obtained as $\{d_2\}$.

The algorithm now collects evidences in support of categorial divergence (d_2):

Table 3 suggests that problematic FT for d_2 is predicative adjunct, i.e., "in dilemma". Thus problematic word is "dilemma". WordNet 2.0 provides only one

sense for dilemma: "*state of uncertainty or perplexity especially as requiring a choice between equally unfavorable option*". Thus the problematic word a_2 is *dilemma#n#1*. A search in PSD_2 for the word that is semantically most similar to a_2 retrieves the entry *motion#n#4* as w_2 and the similarity score $sim(a_2, w_2)$ is computed to be 0.578.

It may be noted that similarity between "dilemma" and "motion" is not apparent at the surface level. However, since in this algorithm the hypernyms of the words concerned are used for computing the similarity value, a positive semantic score has been obtained because the last abstraction level in the hypernyms of "dilemma" and "motion" are same which is "$\Longrightarrow$ state".

Since $0.5 \leq \text{sim}(a_2, w_2) < 1$, the Step 2 of the algorithm suggests that NSD_2 has to be checked for further evidence. From NSD_2, the word w_2' most similar to a_2 is determined, and it is found to be *confusion#n#2* with $\text{sim}(a_2, w_2') = 0.960$. The algorithm therefore determines $s(d_2)$, $s(n_2)$, $m(d_2)$ and $m(n_2)$(see case3c). These values are found to be 0.086, 0.035, 0.332 and 0.497, respectively. Since $m(n_2) > m(d_2)$, Flag_2 is set to 0.

Using step 3 the algorithm now constructs the set D' consisting of divergence types d_i for which the Flags have been set to 1. Evidently, D' is found to be empty. Thus the algorithm suggests that the above input sentence does not give any divergence upon translation to Hindi.

It may be noted that the above decision made by the algorithm is a correct one.

4.3 Illustration 3

Now consider the sentence: My house faces east.

Its parsed version is: @GN> PRON PERS GEN SG1 "i" , @SUBJ N SG "house", @+FMAINV V PRES "face", @OBJ N SG "east" <$.>

Note that the main verb of the input sentence is "face" which is not "be" or "have". Further, the sentence has a subject "my house" and an object "east". Thus the FT-features of the given input sentence are: f_3, f_4 and f_5.

According to the Relevance Table the set D is constructed and it has three elements:

- d_1 i.e. structural divergence
- d_3 i.e. conflational divergence because of sub-types 1 and 2.
- d_4 i.e. demotional divergence due to sub-types 1, 3 and 4.

Evidences for structural divergence (d_1):

The problematic FT for d_1 is the main verb which is "face". Nine senses are provided by WordNet 2.0 for the verb "face" of which sense 3 (*be oriented in a certain direction, often with respect to another reference point; be opposite to*) is the relevant one in this case. Thus problematic word a_1 is "*face#v#3*". From PSD_1 the word w_1 that is most similar to a_1 , is retrieved. Note that w_1 is obtained as *attend#v#1*, and the similarity score, $sim(a_1, w_1)$ is calculated to be 0.660. Since $0.5 \leq sim(a_1, w_1) < 1$, the algorithm now checks the NSD_1. From NSD_1, W'_1 is constructed and w'_1 is found to be *cap#v#1* with $sim(a_1, w'_1) = 0.889$. In this case, the algorithm has to determine $s(d_1)$ and $s(n_1)$. These are found to be 0.444 and 0.151 respectively. Thus, $m(d_1) = \frac{1}{2}*(sim(a_1,w_1)+s(d_1)) = 0.552$ and $m(n_1) = \frac{1}{2}*(sim(a_1,w'_1) + s(n_1)) = 0.520$. Since $m(d_1) > m(n_1)$, the algorithm set $Flag_1$ to be 1.

Evidences for conflational divergence d_3:

The problematic FT for d_3 is also main verb (See Table 3), and therefore the problematic word (a_3) here too is "*face#v#3*". From PSD_3 the word w_3 that is most similar to a_3 is retrieved. In this case the same word *face#v#3* exists

divergence type	$(\mathbf{d}_i)$	$\mathbf{s(d}_i\mathbf{)}$	$\mathbf{m(d}_i\mathbf{)}$
structural	(d_1)	0.444	0.552
conflational	(d_3)	0.086	0.543
demotional	(d_4)	0.204	0.602

Table 6: Values of $s(d_i)$ and $m(d_i)$ for illustration 3

in PSD_3, and therefore $sim(a_3, w_3) = 1.0$. Therefore, due to case 2a $Flag_3$ is set to 1.

Evidences for demotional divergence d_4:

Problematic word a_4 for d_4 is also "*face#v#3*", which too exists in PSD_4. Hence $Flag_4$ is also set to 1.

In Step 3, the set $D' = \{d_1, d_3, d_4\}$ is constructed . The set of possible combinations C (see case 4c) is found to be $\{\{d_1, d_3\}, \{d_3, d_4\}\}$. For a final decision the algorithm now computes the values of $s(d_i)$ and $m(d_i)$ (see case 3c). These values are given in Table6. Using the values given therein the algorithm computes $1/2*(m(d_1) + m(d_3)) = 0.548$ and $1/2*(m(d_3) + m(d_4)) = 0.673$.

Since the latter one is maximum, the algorithm suggests that the above input sentence will have divergence d_3 and d_4 upon translation to Hindi. The above decision of the algorithm is also correct.

Tables 7 and 8 provide few more examples with brief explanation. The overall analysis of each example sentence requires 17 columns. Table 7 contains the column numbers (i) to $(viii)$, and Table 8 contains the column numbers (ix) to $(xvii)$. For ease of understanding one column corresponding to serial number (S. No.) and column number (ii) are given in both the tables. In these Tables, "NA" is used when particular condition is *not applicable* and "Nil" implies that *no word having semantic similarity score greater than 0* has been found in the PSD/NSD.

Table 7: Some Illustrations

S. No	**Sentence**	D	**Problematic Word,** a_i	**Most similar word,** w_i	***sim***(a_i, w_i)	**Is** w_i **a coordinate term?**	**Most similar word,** w_i'	***sim***(a_i, w_i')
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	$(viii)$
1.	She will resolve this issue.	d_1	resolve#v#6	calculate#v#1	0.984	No	resolve#v#6	1.0
		d_3	resolve#v#6	Nil	0.0	NA	NA	NA
		d_4	resolve#v#6	Nil	0.0	NA	NA	NA
2.	I will attend this meeting.	d_1	attend#v#1	attend#v#1	1.0	NA	NA	NA
		d_3	attend#v#1	look#v#5	0.66	No	ride#v#9	0.66
		d_4	attend#v#1	face#v#4	0.75	Yes	NA	NA
3.	This exercise will hurt your back.	d_1	hurt#v#2	trample#v#2	0.96	No	twist#v#9	0.96
		d_3	hurt#v#2	knife#v#1	0.96	No	twist#v#9	0.96
		d_4	hurt#v#2	Nil	0.0	NA	NA	NA
4.	John stabbed Mary.	d_1	stab#v#1	stab#v#1	1.0	NA	NA	NA
		d_3	stab#v#1	stab#v#1	1.0	NA	NA	NA
		d_4	stab#v#1	Nil	0.0	NA	NA	NA
5.	This dish tastes good.	d_3	taste#v#1	taste#v#1	1.0	NA	NA	NA
		d_6	good#a#1	Nil	0.0	NA	NA	NA
6.	This table weighs 100kg.	d_1	weigh#v#1	encounter#v#3	0.660	No	stay#v#1	0.660
		d_3	weigh#v#1	measure#v#3	0.972	No	look#v#3	0.660
		d_4	weigh#v#1	suffer#v#6	0.660	No	look#v#3	0.660
7.	It is windy	d_2	windy#a#1	stormy#a#1	0.75	No	Nil	0.0

Continued ...

Table 7: (continued)

S. No	**Sentence**	D	**Problematic Word,** a_i	**Most similar word,** w_i	***sim***(a_i, w_i)	**Is** w_i **a coordinate term?**	**Most similar word,** w_i'	***sim***(a_i, w_i')
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	$(viii)$
	today.	d_5	windy#a#1	stormy#a#1	0.75	No	Nil	0.0
8.	It will be morning soon.	d_2	morning#n#3	pain#n#2	0.406	No	NA	NA
		d_5	morning#n#3	morning#n#3	1.0	NA	NA	NA
9.	She is in pain.	d_2	pain#n#1	pain#n#2	0.438	No	NA	NA
10.	It suffices.	d_3	suffice#v#1	resemble#v#1	0.782	No	meet#v#5	0.96
		d_4	suffice#v#1	suffice#v#1	1.0	NA	NA	NA
		d_5	suffice#v#1	Nil	0.0	NA	NA	NA
The End								

Table 8: Continuation of Table 7

S. No	D (ii)	$s(d_i)$ (ix)	$s(n_i)$ (x)	$m(d_i)$ (xi)	$m(n_i)$ (xii)	**Flag**$_i$ $(xiii)$	D' (xiv)	C (xv)	$\frac{1}{2}(m(d_i) + m(d_j))$ (xvi)	**Result** $(xvii)$
1.	d_1	NA	NA	NA	NA	0	ϕ	NA	NA	Normal
	d_3	NA	NA	NA	NA	0				
	d_4	NA	NA	NA	NA	0				
2.	d_1	NA	NA	NA	NA	1	d_1	NA	NA	d_1
	d_3	0.075	0.141	0.368	0.401	0				
	d_4	NA	NA	NA	NA	0				
3.	d_1	0.241	0.142	0.601	0.551	1	d_1	NA	NA	d_1
	d_3	0.08	0.145	0.52	0.553	0				
	d_4	NA	NA	NA	NA	0				
4.	d_1	NA	NA	NA	NA	1	d_1, d_3	{d_1,d_3}	NA	d_1, d_3
	d_3	NA	NA	NA	NA	1				
	d_4	NA	NA	NA	NA	0				
5.	d_3	NA	NA	NA	NA	1	d_3	NA	NA	d_3
	d_6	NA	NA	NA	NA	0				
6.	d_1	0.224	0.231	0.442	0.495	0	d_3	NA	NA	d_3; No decision about d_4.
	d_3	0.186	0.219	0.579	0.439	1				
	d_4	0.287	0.287	0.473	0.473	½				
7.	d_2	NA	NA	NA	NA	1	d_2, d_5	{d_2, d_5}	NA	d_2, d_5
	d_5	NA	NA	NA	NA	1				
8.	d_2	NA	NA	NA	NA	0	d_5	NA	NA	d_5

Continued ...

Table 8: (continued)

S. No	D (ii)	$s(d_i)$ (ix)	$s(n_i)$ (x)	$m(d_i)$ (xi)	$m(n_i)$ (xii)	**Flag**$_i$ $(xiii)$	D' (xiv)	C (xv)	$\frac{1}{2}(m(d_i) + m(d_j))$ (xvi)	**Result** $(xvii)$
	d_5	NA	NA	NA	NA	1				
9.	d_2	NA	NA	NA	NA	NA	NA	NA	NA	Normal as $sim(a_2, w_2) < 0.5$, wrong decision
10.	d_4	NA	NA	NA	NA	1	d_4	NA	NA	d_4
	d_5	NA	NA	NA	NA	0				
The End										

4.4 experimental results

In order to evaluate the performance we have used the above algorithm on randomly selected 300 sentences, that are not currently present in our example base. Manual analysis of the translations of these 300 sentences revealed that 32 of them will involve some type of divergence when translated from English to Hindi. Remaining 268 sentences have normal translations.

The output of the algorithm is as follows: It recognized 36 of the sentences to have divergence upon translation, and 261 to have normal translation. For 3 sentences the algorithm could not make any decision. Table 9 summarizes the overall outcome.

	Divergence	Normal
Number of examples	32	268
Experimental results	36	261
Correct results	30	260
Recall %	83.33%	99.62%
Precision %	93.75%	97.39%

Table 9: Results of Our Experiments

The very high value (above 90%) for precision establishes the efficiency of the algorithm in detecting possible occurrence of divergence even before the actual translation is carried out.

There are few examples when the algorithm failed to produce the correct decision. These may be put into three categories:

1. Translation of the input sentence actually involves divergence but the algorithm predicts normal translation. Table 9 indicates that there is one such case in our experiments. Although the algorithm suggests that 261 sentences will be translated normally it has been found that actually 260 of them are correct decisions.

2. The input sentence actually has normal translation but the algorithm predicts divergence. In the experiments carried out by us, we found six such examples. While the algorithm suggests that 36 sentences will involve some type of divergence only 30 of them are correct decisions (see Table 9).

3. The algorithm is unable to decide the nature of the translation of the input sentence. Out of 300 examples tried the algorithm could provide decisions for only 297 (36+261) sentences. For the remaining three sentences the algorithm could not arrive at any decision regarding whether they will be translated normally, or their translations will involve some type of divergence. These are the situations that fall under case 3c of the algorithm.

   Table 7 provides one of the example of this type. Here the input sentence and its translation are: "this table weighs 100kg" $\sim$ (*iss*) *this mez kaa vajan* (weight of this table) *100 kilo* (100 Kg) *hai* (is)". This example has demotional divergence, i.e., d_4. However the algorithm could not give any decision regarding occurence/non-occurrence of d_4 since the values of both $m(d_4)$and $m(n_4)$ are computed to be 0.473.

The algorithm is not able to give correct result in first two cases. We feel that the possible reasons behind the incorrect decisions taken by the algorithm are the following:

- *Lack of robust PSD and NSD.* The present size of the PSD and NSD are 468 and 4132 respectively. Evidently, these numbers are not large enough to deal with all different sentences. As more examples (particularly, those involving divergence) are collected, both the PSD and NSD may be enriched with additional entries. This will in turn enable the algorithm to

measure semantic similarity in a more direct way. As a consequence the number of erroneous decisions will reduce.

- *The value of threshold.* For our experiments we have used 0.5 to be the value of the threshold t. This value has been obtained by carrying out a number of experiments on our example base. However, with more examples this value of t may have to be reassigned, which may in turn improve the quality of the results. Further experiments with more examples need to be carried out to arrive at an optimal value of the threshold t.

5 Conclusion

Occurrence of divergence poses great hindrance in efficient adaptation of retrieved sentences in an EBMT system. This can be dealt with efficiently provided an EBMT system is capable of making a priori decision regarding whether an input sentence will cause any divergence upon translation. This will enable the EBMT system to retrieve a past example more judiciously. However, the primary difficulty in handling divergence is that their occurrences are not governed by any linguistic rules. Hence no straightforward method exists for determining whether a source language sentence will involve any divergence upon translation. The present work is aimed at bridging this gap. This work proposes that an a priori decision may be made in this regard by seeking evidences from the existing example base. In order to achieve the above goal we first analyzed different divergence examples to ascertain the root cause behind occurrence of a divergence. We found that each divergence type can be associated with some Functional Tag (FT) that is instrumental for causing this type of divergence. We call it the "problematic FT" corresponding to that particular divergence. In fact, a detailed analysis of a large number of translation examples revealed

that occurrence of each type of divergence invariably demands certain patterns in the structure of the input sentence. While the presence of certain FTs (including the problematic FT) in the input sentence is mandatory, some other FT features should necessarily be absent in order that the particular divergence type can occur.

However, since divergence in an occasional phenomenon, it is not true that any sentence having the structure required by a particular divergence will certainly involve divergence upon translation. Occurrence of divergence also depends upon semantics of some constituent words. To measure the semantic similarity between words two dictionaries, viz. "problematic sense dictionary" (PSD) and "normal sense dictionary" (NSD), have been created.

Given an input, these knowledge bases are referred to seek evidence in support/against divergence. Evidences used are of the following types:

(a) The Functional Tags of the constituent words of a given input;

(b) Semantic similarity of these constituent words with words in the PSD and NSD;

(c) Frequency of occurrence of different divergence types in the example base; and

(d) Which divergence types may co-occur in the translation of an input sentence?

Since divergence is a language-dependent phenomenon we have chosen specific languages viz. English and Hindi as the source and target languages for this work. However, due to overwhelming similarities of various Indian languages, such as Bangla, Gujrati, Marathi, with Hindi, we feel that similar scheme should work with respect to translations from English to these languages as well. Hence

the scheme presented here should find significant applications in linguistic research/projects involving English and other Indian languages (such as EMILLE : http: //www.emille.lancs.ac.uk/home.htm, Example-based Machine Translation system (Shiva): (http://ebmt.serc.iisc.ernet.in/mt/login.html)).

Extension of this work to European languages may, however, require some additional work. Although study of divergence on English to Hindi translation ([3],[5]) finds its root in the study of divergence between European languages [4], it has been observed that all the divergence types given therein do not apply with respect to Indian languages. Further, definitions of some of the divergence types had to be modified to suit the requirements of Hindi. A systematic analysis of divergence and normal translation examples needs to be carried out, and appropriate sense dictionaries need to be created in order to develop any such schemes for dealing with translation divergence between European languages. The work presented here should provide the required guidelines for any such studies.

The experiments carried out by us resulted in very high values of precision and recall. However, more experiments need to be done to establish this scheme as a key technique for dealing with divergences for an EBMT system.

One may envisage the following shortcomings in the scheme presented here:

1) The algorithm in its present form cannot deal with "possessional" divergence that has been defined in English to Hindi context [7]. We have observed that occurrence of possessional divergence depends not on a single problematic FT. Rather study of different features (e.g. subject, object, their pre-modifiers and their hypernyms) of the input sentence is needed to arrive at any conclusion in this regard. The scheme proposed here needs to be further extended to deal with multiple problematic FTs in order to make any prior decisions regarding occurrence/non-occurrence of

possessional divergence in the translation of an input sentence.

2) Creation of the sense dictionaries is an important background work required for implementation of the proposed scheme. The sense dictionaries (PSD and NSD) used in this work have been created manually. Some suitable Word Sense Disambiguation techniques may have to be developed/used to accomplish this task.

3) At present the decisions made by the scheme concerns with divergence types only. We feel that the scheme may be further extended to deal with various sub-types that are associated with each divergence type. More examples involving each of these sub-types need to be obtained and analyzed for any such extension. Our present example base does not have sufficient number of examples for each sub-type. We are currently working on acquisition of such examples for possible extension of the algorithm in this direction, and also to improve upon the performance of the present scheme.

Appendix A

Divergence Definitions and Their Sub-types

Divergence Type	**Sub-Types**	**Necessary Properties for the English sentence (E)**	**Changes in sentence structure in English (E) to Hindi (H)**
Structural	-	The root form of the main verb should not be 'be" or "have". An object should be present, and it has to be a noun phrase (NP).	The object of E is realized as an object in H, and it has to be a prepositional phrase (PP).
Categorial	Sub-type1	The root form of the main verb should be "be". The sentence should have a SC which should be an adjective.	The subjective complement (SC) of E becomes the main verb of H and its root form should not be "*ho*"(i.e. "be" in H
	Sub-type2	The root form of the main verb should be "be". The sentence should have a SC which is a NP.	The SC of E is realized as the main verb of H and its root form should be other than "be"
Continued ...			

(continued)

Divergence Type	**Sub-Types**	**Necessary Properties for the English sentence (E)**	**Changes in sentence structure from English (E) to Hindi (H)**
Categorial	Sub-type3	The root form of the main verb should be "be". The sentence must have a predicative adjunct (PA), and it has to be an adverb.	The PA of E becomes the main verb of H, and its root form is other than "be".
	Sub-type4	The root form of the main verb should be "be". Further, a PA should be present, and it should be a PP.	The PA of E becomes the main verb of H, and its root form should not be "be"
Conflational	Sub-type1	The root form of the main verb should be other than "be"/"have".	The main verb of E adds some extra information in H in the form of PP.
	Sub-type2	The root form of the main verb should be other than "be"/"have". The sentence should have a Subject.	The main verb of E provides the subject of H. The subject of E is realized as a postmodifier of the new subject, and it should be a PP.
Continued ...			

(continued)

Divergence Type	**Sub-Types**	**Necessary Properties for the English sentence (E)**	**Changes in sentence structure from English (E) to Hindi (H)**
Demotional	Sub-type1	The root form of the main verb should be other than "be"/"have". Also, the sentence should have an object.	The main verb and object of E are realized as Predicative adjunct (PA) in H.
	Sub-type2	The root form of the main verb should be other than "be"/"have". There should be a verb complement(VC) in the sentence	The main verb and VC of E are realized as a PA in H.
Demotional	Sub-type3	The root form of the main verb should be other than "be"/"have". Also, the sentence should have an object.	The main verb of E is realized as an adjectival SC in H. Further, the object of E becomes an adjective complementation, which is a PP, in H
Continued ...			

(continued)

Divergence Type	**Sub-Types**	**Necessary Properties for the English sentence (E)**	**Changes in sentence structure from English (E) to Hindi (H)**
	Sub-type4	The root form of the main verb should be other than "be"/"have". Also, the sentence should have an object.	The main verb of E becomes an adjectival SC. The object and the subject of E are realized as the subject and the postmodifier of SC, respectively.
Pronominal	Sub-type1	The Subject of the sentence has to be "It". The root form of the main verb should be "be". Further, an SC should be present and it has to be an NP	The SC of E is realized as the subject of H.
	Sub-type2	The Subject of the sentence has to be "It". The root form of the main verb should be "be". Further, an SC should be present and it has to be an adjective.	The SC of E is realized as the subject of H.
Continued ...			

(continued)

Divergence Type	**Sub-Types**	**Necessary Properties for the English sentence (E)**	**Changes in sentence structure from English (E) to Hindi (H)**
Pronominal	Sub-type3	The Subject of the sentence has to be "It". The root form of the main verb should be "be". Further, to-infinitive form of some verb should also be present.	The to-infinitive form in E is realized as the subject of H.
	Sub-type4	The Subject of the sentence has to be "It". The root form of the main verb should be other than "be"/"have".	The main verb of E is realized as the subject of H.
Nominal d_6	Sub-type1	The root form of the main verb should be other than "be"/"have". The sentence should have an adjectival SC.	The SC of E becomes the subject of H. The subject of E is realized as the object of H.
Continued ...			

(continued)

Divergence Type	**Sub-Types**	**Necessary Properties for the English sentence (E)**	**Changes in sentence structure from English (E) to Hindi (H)**
	Sub-type2	The root form of the main verb should be other than "be"/"have". The sentence should have an adjectival SC.	The SC of E is realized as the subject of H. The subject of E becomes the Verb complement (VC) in H.

Appendix-B

In this work we have used the ENGCG parser (http://www.lingsoft.fi/cgi-bin/engcg) for parsing the English sentence. Most of the FTs that are relevant for this work are obatined directly from the parser. Description of these FTs are given below:

@+FAUXV – Finite Auxiliary Predicator
(e.g. He *can* read.)

@-FAUXV – Nonfinite Auxiliary Predicator
(e.g. She may *have* read.)

@+FMAINV – Finite Main Predicator
(e.g. He *reads*.)

@-FMAINV – Nonfinite Main Predicator
(e.g. She has *read*.)

@SUBJ – Subject
(e.g. *He* reads.)

@OBJ – Object
(e.g. She read a *book*.)

@PCOMPL-S – Subject Complement
(e.g. He is a *fool*.)

@ADVL – Adverbial
(e.g. She came home *late*. She is *in* the car.)

@DN> – Determiner

(e.g. He read *the* book.)

@AN> – Premodifying Adjective

(e.g. The *blue* car is mine.)

@QN> – Premodifying Quantifier

(e.g. He had *two* sandwiches and *some* coffee.)

@GN> – Premodifying Genitive

(e.g. *My* car and *Bill's* bike are blue.)

@<P – Other Complement of Preposition

(e.g. He is in the *car*.)

Each FT tag is prefixed by "**@**" in contradistinction to other types of tags. Some tags include an angle bracket, "<" or ">". The angle bracket indicates the direction where the head of the word is to be found.

Some of the functional tags that are required for divergence identification algorithms are not directly given by the available parsers. These FTs are Adjunct (A), predicative adjunct (PA) and VC (verb complement) We have formulated rules for obtaining these FTs by using information available in the morpho tags of the underline sentence.

Appendix-C

Semantic similarity between two words is computed on the basis of their semantic distance (sd), as follows:

$sim(a,b) = 1 - (sd(a,b))^2$

The semantic similarity score lies between 0 and 1. Semantic distance [Stetina et. al., 1998] between two words, say a and b, is computed as:

- *Semantic Distance for Nouns and Verbs*

$$sd(a,b) = \frac{1}{2}\left(\frac{H_a - H}{H_a} + \frac{H_b - H}{H_b}\right)$$

Here, H_a is the depth of the hypernyms of a, H_b is the depth of the hypernyms of b, and H is the depth of their nearest common ancestor.

- *Semantic Distance for Adjectives and Adverb*

  $sd(a,b) = 0$ for the same adjectival synsets (including Synonymy)

  $sd(a,b) = 0$ for the synsets in antonym relation $ant(a,b)$

  $sd(a,b) = 0.5$ for the same synsets in the same similarity cluster and antonym relation ant(a,b)

  $sd(a,b) = 1$ for all other synsets.

References

[1] Brown, R. D. Example-Based Machine Translation in the Pangloss System. In *Proceeding of COLING-96:*, pages 169 – 174, Copenhagen, 1996.

[2] Carl, M. and A. Way. Advances in Example-Based Machine Translation Series : Text, Speech and Language Technology, Vol 21. Springer, 2003.

[3] Dave, S., Jignashu Parikh and Pushpak Bhattachary. Interlingua Based English Hindi Machine Translation and Language Divergence. *Journal of Machine Translation (JMT)*, 17:, September, 2002.

[4] Dorr, B. J. Machine Translation: A View from the Lexicon. MIT Press, Cambridge, MA, 1993.

[5] Gupta D. and N. Chatterjee. Divergence in English to Hindi Translation: Some Studies. *International Journal of Translation*, 15:5–24, 2003a.

[6] Gupta D. and N. Chatterjee. Identification of Divergence for English to Hindi EBMT. In *Proceedings of the MT SUMMIT IX*, pages 141-148, New Orleans, LA, September 2003b.

[7] Gupta D. and N. Chatterjee. An FT and SPAC Based Divergence Identification Technique for English to Hindi EBMT. Submitted to Machine Translation, Kluwer Academic Publishers. Kluwer Academic Publishers, 2004.

[8] Nagao M. A Framework of a Mechanical Translation Between Japanese and English by Analogy Principle. In em A. Elithorn and R. Banerji(eds), Artificial and Human Intelligence,Pages 173-180, Amsterdam: North-Holland 1984.

[9] Stetina J., S. Kurohashi, and M. Nagao. General Word Sense Disambiguation Method Based on a Full Sentential Context. In *Proceedings of COLING-ACL Workshop*, Montreal, Canada, 1998.