Divergence patterns between English and Sanskrit Machine Translation

Vimal Mishra[1]
R.B.Mishra[2]

[1] Research Scholar, Department of Computer Engineering, Institute of Technology, Banaras Hindu University, (IT-BHU), Varanasi-221005, U.P, India
[2] Professor, Department of Computer Engineering, Institute of Technology, Banaras Hindu University, (IT-BHU), Varanasi-221005, U.P., India
[1] vimal.mishra.upte@gmail.com
[2] ravibm@bhu.ac.in

Abstract. The translation divergence is a challenging problem in the area of machine translation. A detailed study of divergence issues in machine translation is required for their proper classification and detection. The language divergences between English and Sanskrit can be considered as representing the divergences between SVO (Subject - Verb - Object) and SOV (Subject - Object - Verb) classes of languages. This topic needs exploration to identify different sources of translation divergence between English and Sanskrit. This paper discusses translation patterns between English and Sanskrit to identifying the potential topics of translation divergences. The typical type (specific to language pair such as English and Sanskrit) of divergence is based on different aspects such as linguistic to socio-and psycho-linguistic, role of conjunctions and particles, participle, gerunds and socio-cultural aspects. We have proposed the detection rules for these types of divergence related sentences and apply the adaptation rules on it. The results of divergence are shown in GUI form. We evaluate the results of our system with different evaluation methods of machine translation.

Keywords: translation patterns, Sanskrit, Machine Translation, translation divergence, divergence pattern.

(Received March 28, 2009 / Accepted August 17, 2009)

1. Introduction

The divergence is a common problem in translation between two natural languages. The language divergence [1] [2] arises when lexically and syntactically similar sentences of the source language do not translate into sentences that are similar in lexical and syntactic structure in the target language.

Depending upon the direction of translation (from English to Sanskrit and vice versa), the divergence may be classified as unilateral or bilateral. If divergence occurs when we translate from English to Sanskrit or vice versa then this type of divergence is fallen into unilateral type of divergence. If divergence occurs when we translate from English to Sanskrit; and Sanskrit to English then this type of divergence is called as bilateral type of divergence. This phenomenon may occur in any pair of languages for machine translation whether it may be English to Spanish and English to German [2], Spanish to English [6], and English to Hindi [1]. Some previous work is described below using different approaches of machine translation for the divergence detection between different pairs of languages.

In the transfer approach of translation divergence, there is a transfer rule for transforming a source language (SL) sentence into target language (TL), by performing lexical and structural manipulations. These transfer rules are formed in several ways such as: (i) With manual encoding [7]; and (ii) With analysis of parsed aligned bilingual corpora [15]. The lexical and structural based divergences are dealt in this approach.

In the Interlingua approach, the identification and resolution of divergence are based on two mappings GLR (Generalized Linking Routine), CSR (Canonical syntactic Realization) and a set of LCS (Lexical Conceptual Structure) parameters. The translation divergence occurs when there is an exception either to GLR or to CSR (or to both) in one language but not in the other. This situation permits one to formally define a classification of all possible lexical-semantic divergences that could arise during translation. This approach has been used in the UNITRON system [2] that pursues translation from English to Spanish and English to German. In Universal Networking Language (UNL) based Interlingua approach, the sentences are represented using hyper graphs with concepts as nodes and relations as directed arcs. A dictionary of UW (Universal Word) is maintained. A divergence is said to occur if the UNL expression generated from the both source and target language analyzer differ in structure. Dave et al [1] has proposed UNL based Interlingua approach for English to Hindi machine translation.

The MATADOR System [6] uses this approach for translation between Spanish and English. In this approach, a symbolic over generation is created for a target glossed syntactic dependency representation of SL sentences which uses rich target language resources, such as word-lexical semantics, categorical variations and sub-categorization frames for generating multiple structural variations. This is constrained by a statistical TL model that accounts for possible translation divergences. Then, a statistical extractor is used for extracting a preferred sentence from the word lattice of possibilities. This approach bypasses explicit identification of divergence, and generates translations (which may include divergence sentences) otherwise.

Each of the above approach has problems when we apply in English to Sanskrit machine translation. For example, GHMT (Generation Heavy Machine Translation) approach requires rich resources for the target language (here, Sanskrit) which is not available for Sanskrit now-a-days. The Interlingua approach requires deep semantic analysis of the sentences and creation of exhaustive set of rules to capture all the lexical and syntactic variation may be problem in English to Sanskrit translation. While in case of UNL based Interlingua approach, each UW of the dictionary contains deep syntactic, semantic and morphological knowledge about the word. Creation of such UW dictionary for a restricted domain is difficult and rarely available. With respect to Sanskrit, the major problem in applying the above approach is that linguistic resources are rarely available in Sanskrit. The Sanskrit is free word order language but the preferred pattern is SOV. We take SOV pattern of Sanskrit to describe translation in English to Sanskrit.

The rest of the work in this paper is divided into following sections. Section 2 describes divergence detection in English to Sanskrit translation which is based on different aspects such as linguistic to socio-and psycho-linguistic, role of conjunctions and particles, participle, gerunds and socio-cultural aspects. Sections 3 represent implementation of English to Sanskrit Machine Translation (EST) system. Section 4 shows results from our EST model. Section 5 shows the evaluation of our EST system with table and column chart. Section 6 gives the conclusions.

2. Divergences and its detection in English to Sanskrit translation

The divergence is a language dependent phenomenon, it is not expected the same set of divergences will occur across all languages. We classify divergences in two types: conventional as well as typical type of convergences. Dorr [3] classifies divergence in seven broad types, which is lexical-semantic divergences for translating among the European languages, such as Structural divergence, Conflational divergence, Categorical divergence, Promotional Divergence, Demotional Divergence, Thematic Divergence and Lexical divergence. We classify divergences as the conventional type which is based on Dorr's classification of translation divergence to examine the different types of translation divergence in English and Sanskrit [9] [10].

In this paper, we describe typical type of divergences in English to Sanskrit machine translation which is based on different aspects such as linguistic to socio-and psycho-linguistic, role of conjunctions and particles, participle, gerunds and socio-cultural aspects. This work presents new sources and topics of translation divergence in English to Sanskrit and Sanskrit to English machine translation. The issue of divergence between a pair of languages is associated by a number of factors such as linguistic to socio- and psycho-linguistic aspects of the languages involved [12] [13] [14]. We discuss translation divergence between English sentence (ES) and Sanskrit sentence (SS) which is based on different aspects that is given below.

2.1. Conjunctions and particles related divergence

We examine another source of divergence between English and Sanskrit which is related to conjunctions and particles (CP) in Sanskrit such as "yat" or "vaa" ("or" in English). The "vaa" is an indeclinable element (particle) in Sanskrit which is used in multiple roles that have multiple mapping patterns in English. The "yat" or "vaa" is mainly used as a sentence complementizer, but can also be used to indicate alternate conjunction in an affirmative sentence and an interrogative sentence in Sanskrit. The English sentence and their Sanskrit translation are given below.

(i) ES: Ram has gone either to Delhi or to Mumbai.

SS: Rama Dilliim gatavaan asti vaa Mumbaim.

(Ram) (to Delhi) (gone) (has) (or)(to Mumbaim)

(ii) ES: Does Ram study or sleep?

SS: Kim Ramah pathati vaa shete.

(Does) (Ram) (study) (or) (to sleep)

In the above example, "vaa" (either-or, or in English) is a coordinate conjunction particle in Sanskrit that joins two clauses or phrases. The above examples show the conjunctions and particles related divergence in Sanskrit as the bilateral type of divergence. For the detection and correction of the conjunctions and particles related divergence, we use Rule Module I which is described below.

Rule Module I

Rule1 If (CP of ES = "either-or" || "or") Then Place "vaa" as the second last word in concatenation of SS.

In the above rule, if-clause denotes detection (condition) and then-clause denotes action (adaptation).

2.2. Participle related divergence

Another source of divergence between English and Sanskrit is related to the participle in Sanskrit. In Sanskrit, the participle is formed by using "tuman" suffix. The English sentence and their Sanskrit translation are given below.

(i) ES: Ram got happy to come.

SS: Ramah aagantum prasannam abhavat.

(Ram) (to come) (happy) (got)

(ii) ES: She wants to go.

SS: Saa gantum ichchati.

(She) (to go) (wants)

The example (i) and (ii) show the participle related divergence as the bilateral type of divergence. For the detection and correction of the participle related divergence, we use Rule Module II which is described below.

Rule Module II

Rule2 If (ES= "to" followed by verb) Then Add "ntum" as suffix in the verb of Sanskrit.

2.3 Gerunds related divergence

Another source of divergence between English and Sanskrit is related to the gerunds in Sanskrit. We classify gerunds such as gerund type1, gerund type2 and gerund type3 when gerund is used as a subject of a verb, gerund is used as an object of a verb and gerund is governed by a preposition "of", respectively. The English sentence and their Sanskrit translation are given below for gerund type1, gerund type2 and gerund type3 in examples (i) to (ii), (iii) to (iv) and (v) to (vi), respectively.

(i) ES: Sleeping is necessary for life.

SS: Svapnam jiivanaaya nitaantam asti.
(sleeping) (for life) (necessary) (is)

(ii) ES: Laughing is not right.
SS: Hasnam na uchitam.
(laughing) (not) (right) (is)

(iii) ES: I like reading book.
SS: Aham pathitum puskatam vanaami.
(I) (reading) (book) (like)

(iv) ES: He enjoys sleeping in room.
SS: Sah svapitum prakoshtam vanaami.
(He) (sleeping) (in room) (enjoys)

(v) ES: I am fond of eating mango.
SS: Aham aamrah bhuktavaa preman asmi.
(I) (mango) (of eating) (fond) (am)

(vii) ES: Having gone to town, Ram drinks water.
SS: Ramah gramam gatvaa jalam pibati.
(Ram) (to town) (having gone) (water) (drinks)

The above examples show the gerunds related divergence as the bilateral type of divergence. For the detection and correction of the gerunds related divergence, we use Rule Module III which is described below.

Rule Module III

Rule3 If (First word of ES= (verb + "ing")) Then Add "nam" as suffix to the Sanskrit verb.

Rule4 If (Second verb of ES= (verb + "ing")) Then Add "itum" as suffix to the Sanskrit verb.

Rule5 If (verb of ES= "of" followed by (verb + "ing")) Then Go to Rule7.

Rule6 If (First word of ES= "having" +3rd form of verb) Then Add "tvaa" as suffix to the Sanskrit verb.

In the case of complex or compound sentence, if gerund related sub-sentence comes then we apply above Rule Module III to translate the English sentence (ES). For example, the sentence "I in English) as subject in Sanskrit and use dhaatu
corresponding to this.

// S1 and S2 are sub sentence of S that conjunct with "and".

assure that sleeping is necessary for life". Our EST system breaks this sentence into two sub-sentences as "I assure" (ES1) and "sleeping is necessary for life" (ES2). Thus, this English sentence takes the form of ES1- Conjunction- ES2. Our EST system translates ES1 as simple type of sentence, put the conjunction in Sanskrit and translate ES2 as gerund type of sentence.

2.4. Social-cultural related divergence

The honorific feature is marked by using the plural of the verb in Sanskrit while English lacks this feature. The English sentence and their Sanskrit translation are given below.

(i) ES: The President has arrived and he will deliver a lecture now.
SS: Rastrapati aagamat tathaa te samprati bhashranam daasyati.
(The President) (has arrived) (and) (they) (deliver) (a lecture) (will)

In both English to Sanskrit and Sanskrit to English machine translation, the social-cultural related divergences arise. For the detection and correction of the social-cultural related divergence, we use Rule Module IV which is described below.

Rule Module IV

Rule7 If (Noun of S1 = HN) Then Use plural number of nominative case of subject in Sanskrit and use dhaatu corresponding to this.

// Honorific Noun (HN) = {President, Prime-Minister, Chief-Minister, Minister, Leader, Teacher, Doctor, Engineer)

Rule8 If (Subject of S2 = "he") Then Use plural number of nominative case of sah ("he" in English) as subject in Sanskrit and use dhaatu corresponding to this.

Rule9 If (Subject of S2 = "she") Then Use plural number of nominative case of saa ("she"

3. Implementation

The flowchart for the implementation of algorithm mentioned in the above section is given in figure 1. The divergence detection module detect the different type of divergences such as conjunctions and particles related, participle related, gerunds related and socio-cultural related divergences. The intermediate translation module is described in figure2. After detection of divergences, we apply adaptation to correct the translation which is described in section 2 for each type of divergences. We get final (correct) translation after applying adaptation on it.

We have developed English to Sanskrit machine translation(EST) model that comprised of combined two approaches: rule based model and the dictionary matching by ANN (Aritificial Neural Network) model. Our EST model has been implemented on windows platform using Java. The ANN model is implemented using MATLAB 7.1 neural network tool. We use feed forward ANN that gives matching of equivalent Sanskrit word of English word which handles noun and verb. The rule based model is generated Sanskrit translation of the given input English sentence using rules that extract verb and noun form for Sanskrit. The different divergence detection algorithms are handled well in our EST model. In this paper, we describe the main module of our EST model that is given below, in brief. We basically perform three steps in ANN based systems which are given as: (1) Encoding of User Data Vector (UDV); (2) Input-Output generation of UDV; and (3) Decoding of UDV. In the Encoding of UDV, we write alphabet (a-z) into five bit binary in which alphabet "a" as 00001, to avoid the problem of divide by zero and alphabet "z" as 11010. For the training into ANN system, we make the alphabet to decimal coded form which is obtained by dividing each to thirty-two.

This gives us input word in decimal coded form and output in corresponding Sanskrit word in roman script. In the Input-Output generation of UDV, we prepare the input-output pair of data for the two to five characters verb and noun in English as input and corresponding verb and noun in Sanskrit as output. After preparing the UDV, we train the UDV through feed forward ANN and then test the UDV. We get the output of Sanskrit word in the UDV form. In the decoding of UDV, each values of a data set is compared with the decimal coded values of alphabet, one by one and the values with minimum difference is taken with its corresponding alphabet.

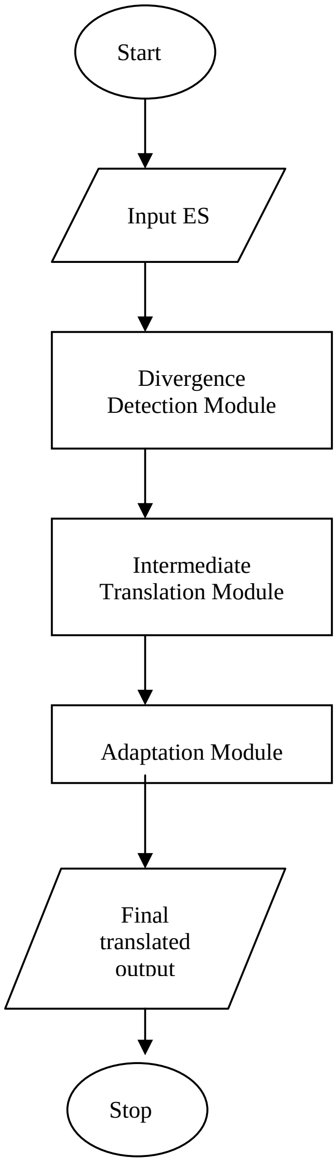


Figure1. A flowchart that shows the detection and the correction of divergences in our EST system

We have a data set of 125 input-output pair for verb. The input, hidden and output values for verb is taken 5, 38 and 6. The training is terminated at a training error of 10^{-3} after 300 epochs. For the noun, we have 100 input-output pair in which the input, hidden and output values are taken 5, 15 and 7. This training is terminated at a training error of 10^{-2} after 300 epochs. Figure 2 shows the information flow in our EST model. We have a data set of 125 input-output pair for verb. The input, hidden and output values for verb is taken 5, 38 and 6. The training is terminated at a training error of 10^{-3} after 300 epochs. For the noun, we have 100 input-output pair in which the input, hidden and output values are taken 5, 15 and 7. This training is terminated at a training error of 10^{-2} after 300 epochs. Figure 2 shows the information flow in our EST model.

3.1. Sentence Tokenizer Module

The sentence tokenizer module split the English sentences into tokens (words) using split method of string tokenizer class in Java. The outputs of the sentence tokenizer module are given to POS Tagger module.

3.2 POS Tagger Module

The POS (Part–of-Speech Tagging) is the process of assigning a part–of-speech (such as a noun, verb, pronoun, preposition, adverb and adjective) to each word in a sentence. In POS Tagger module, the Part–of-Speech (POS) tagging is done on each word in the input English sentence. The output of POS tagger module is given to rule base engine.

3.3. GNP detection Module

The GNP detection module detects the gender, number and person of the noun in English sentence. The English language has three genders: masculine, feminine and neuter; two numbers: singular and plural and three persons: first, second and third.

3.4. Tense and Sentence detection Module

The English has three tenses: present, past and future; and four forms of each tense such as indefinite, continuous, perfect and perfect continuous. The tense of English sentence is determined by using rules. The sentence detection gives the structure, form and type of sentence.

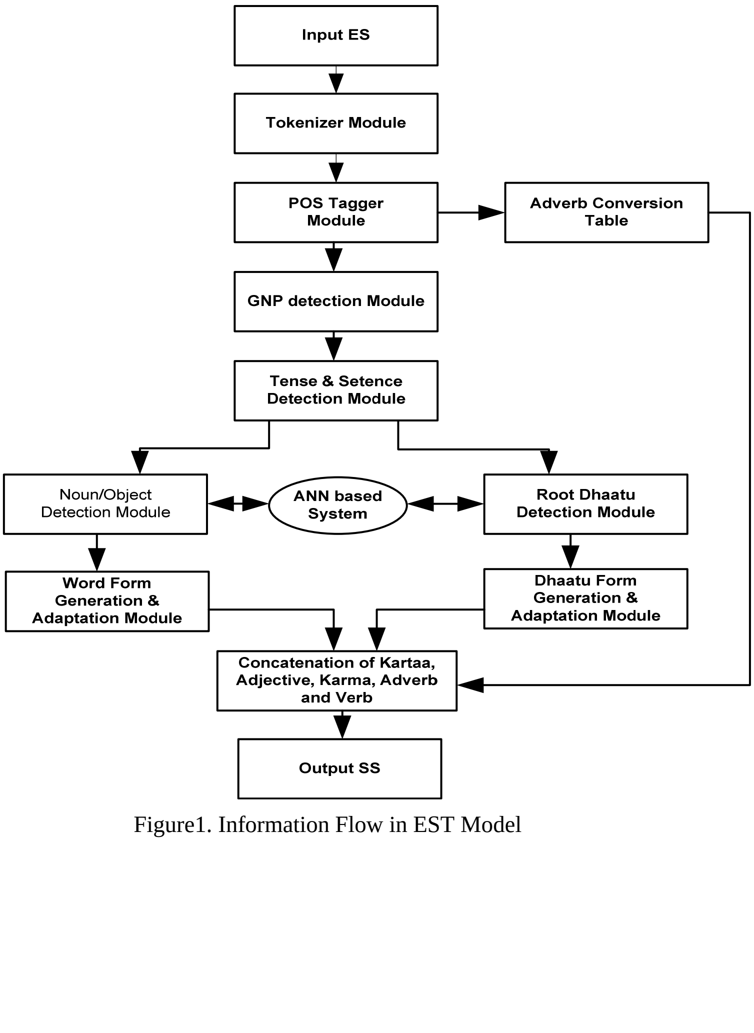


Figure1. Information Flow in EST Model

3.5. Noun and object detection Module

This module gives noun for Sanskrit of the equivalent English noun. It uses ANN method for the selection of noun for Sanskrit. The adaptation rules are used to generate the word form.

3.6. Root Dhaatu detection Module

This module gives verb for Sanskrit of the equivalent English verb. It uses ANN method for the selection of verb for Sanskrit. We apply adaptation rules to generate the required dhaatu form.

3.7. Adverb Conversion Table

This table has the most used adverb for Sanskrit of the equivalent English adverb, which are stored in a database and having one to one correspondence in the table I.

S. No.	Adverb in English	Adverb in Sanskrit
1	Slowly	Shanaih
2	Suddenly	Akasmaat
3	Everyday	Sarvatra
4	Continuous	Anisham
5	Fast	Durtah
6	Always	Sadaa
7	Today	Adah
8	Daily	Pratidinam

Table I. Commonly used adverb in English and corresponding in Sanskrit.

4. Results

Our EST system handles most of the divergences from English to Sanskrit machine translation. It is evident that in general, it is not possible to deal with all kinds of divergence in this paper. But we have covered the most of the commonly found divergence.

Our EST system is ANN and rule based model. During the development of EST model, an attempt is made to device rules that are based on algorithm of translation divergence detection. The output of our EST system for conjunctions and particles related divergence, participle related divergence and gerunds related divergence in EST is shown in figure 3, 4, and 5, respectively.

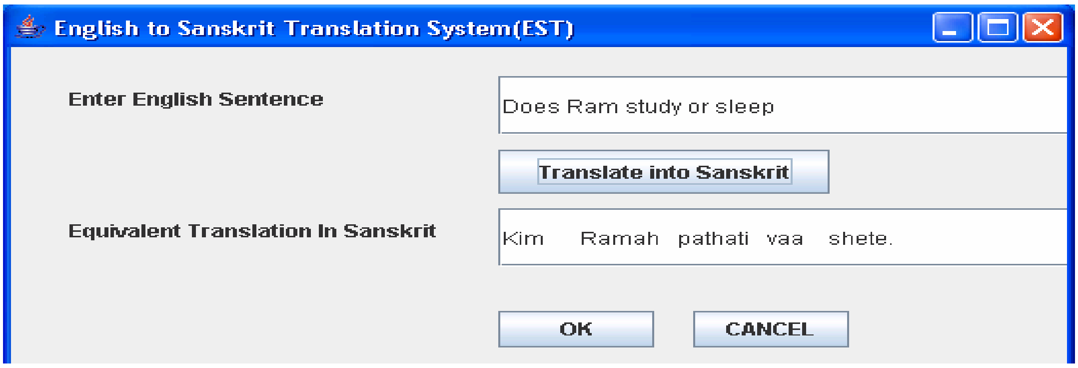


Figure3. Handling of conjunctions and particles related divergence in EST.

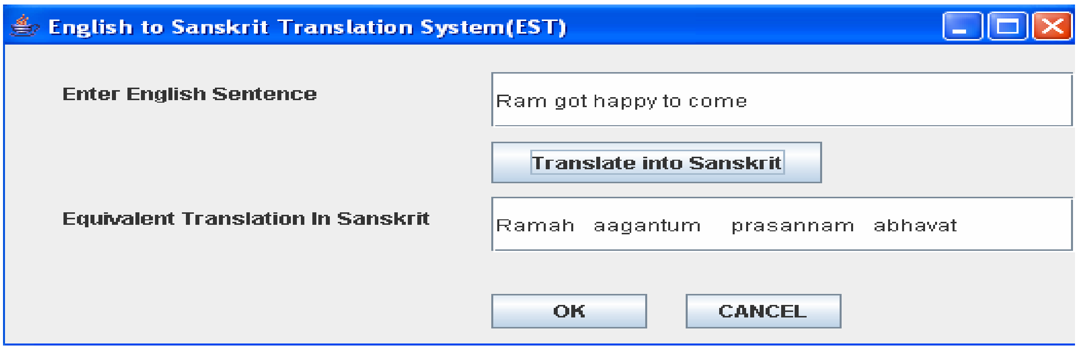


Figure4. Handling of participle related divergence in EST.

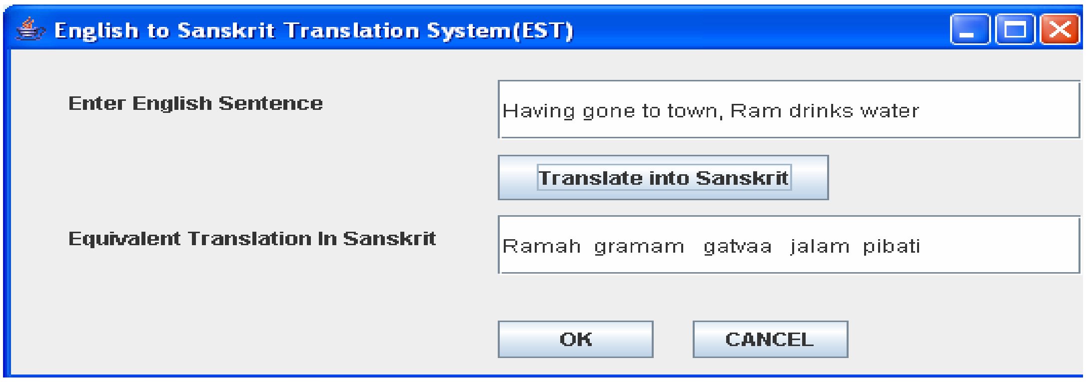


Figure5. Handling of gerund related divergence in EST.

5. Evaluation

We evaluate the performance of our EST system that handle the divergences of conjunctions and particles, participle, gerunds and socio-cultural aspects related of sentences using different MT evaluation methods such as BLEU (Bi**L**ingual Evaluation Understudy) [11], unigram Precision (P), unigram Recall (R), F-measure (F) [8] and METEOR (M) [12] score. The evaluation scores of our EST system are encouraging which are calculated among randomly 10 selected sentences (ES) with our EST system (C) including reference translations (R), that are given below.

1. ES: Does Ram study or sleep?

   C: Kim Raamah pathati vaa shete.

   R: Kim Raamah adhiite vaa shete.

2. ES: Ram got happy to come.

   C: Raamah aagantum prasannam abhavat.

   R: Raamah aagantum muditam abhavat.

3. ES: Having gone to town, Ram drinks water.

   C: Raamah gramam gatvaa jalam pibati.

   R: Raamah gramam itvaa jalam pibati.

4. ES: I like reading book.

   C: Aham pathitum pustakam vanaami.

   R: Aham pathitum pustakam ichchati.

5. ES: The President has arrived and he will deliver a lecture now.

   C: Rastrapati aagamat tathaa te samprati bhashranam daasyati.

   R: Rastrapati aagamat te samprati bhashranam daasyati ca.

6. ES: Ram has gone either to Delhi or to Mumbai.

C: Rama Dilliim gatavaan asti vaa Mumbaim.

R: Rama Dilliim gatavaan asti yat Mumbaim.

7. ES: She wants to go.

C: Saa gantum ichchati.

R: Saa gantum ichchati.

8. ES: Sleeping is necessary for life.

C: Svapnam jiivanaaya nitaantam asti.

R: Svapnam jiivanaaya pradhanam asti.

9. ES: Laughing is not right.

C: Hasnam na uchitam.

R: Hasnam nochitam.

10. ES: I am fond of eating mango.

C: Aham aamrah bhuktavaa preman asmi.

R: Aham aamrah khaadatavaa preman asmi.

The evaluation scores for randomly ten selected divergence type of sentences are shown in table II.

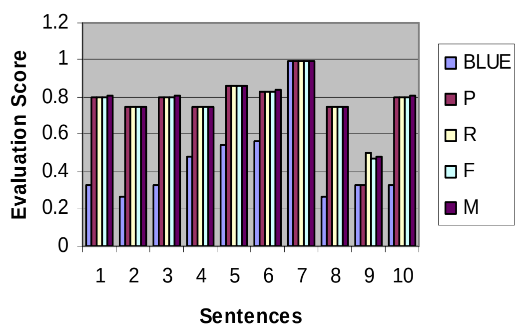
S	BLUE	P	R	F	M
1	0.325	0.8	0.8	0.8	0.811
2	0.271	0.75	0.75	0.75	0.7516
3	0.325	0.8	0.8	0.8	0.811
4	0.48	0.75	0.75	0.75	0.7516
5	0.544	0.8571	0.8571	0.8571	0.8596
6	0.567	0.833	0.833	0.833	0.8387
7	1.0	1.0	1.0	1.0	1.0
8	0.271	0.75	0.75	0.75	0.7516
9	0.333	0.333	0.5	0.4766	0.4798
10	0.325	0.8	0.8	0.8	0.811

Table II. Performance evaluation scores for divergence type of sentences.

The comparative score of different MT evaluation methods such as BLEU (Bi**L**ingual Evaluation Understudy), unigram Precision (P), unigram Recall (R), F-measure (F) and METEOR (M) are shown in figure6.

The performance of work can be evaluated with two perspectives: one is the effectiveness of the computing method with its inherent characteristics and the other is from the linguistic perceptiveness. From the computational perspective, the inherent characteristics of RBS such as modularity representation of facts and ANN model characteristics, optimal search strategy have been effectively used for the translation.

Figure 6. Performance evaluation score for divergence related sentences.



6. Conclusions

The Hindi language is originated from the Sanskrit language. Our paper describes the problem of translation divergence in English to Sanskrit machine translation in view of some existing similar work with respect to divergence from English to Hindi machine translation [1] [4] [5] [13] . But, these researchers have presented only the linguistic view and they have mentioned in their work neither the computing algorithm nor the implementation strategy. We are performed a novel method that uses rules and ANN technique to detect and implement the adaptation rules for the divergence in English to Sanskrit machine translation. The work in this paper is the first work of translation divergence in English to Sanskrit translation. Our future work is carrying to perform case based reasoning in combination of rule based and ANN model for this purpose.

References

[1] Dave, Sachi, Parikh, Jignashu, Pushpak Bhattacharya., *Interlingua-based English–Hindi Machine Translation and Language Divergence,* Machine Translation, Vol. 16, pp. 251–304, Kluwer Academic Publishers, 2001.

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

[3] Dorr, B. J., *Machine Translation Divergences: A Formal Description and Proposed Solution*, Association for Computational Linguistics, pp 597-633, 1994.

[4] Gupta, Deepa and Chatterjee, N., *Study of Divergence for Example Based English-Hindi Machine Translation*, Proceeding of the STRANS 2001, Indian Institiute of Technology Kanpur, pp.132-139, 2001.

[5] Gupta, Deepa and Chatterjee, N**.,** *Identification of Divergence for English to Hindi EBMT*, Proceeding of the MT SUMMIT IX, New Orleans, LA, pp. 141-148, 2003.

[6] Habash, N., Generation-Heavy Hybrid Machine Translation, PhD thesis, University of Maryland, College Park, 2003.

[7] Han, C., Benoit, L., Martha, P., Owen, R., Kittredge, R., Korelsky, T., Kim, N. and Kim, M., Handling structural divergences and recovering dropped arguments in a Korean/English machine translation system, Proceedings of the Fourth Conference of the Association for Machine Translation in the Americas, AMTA-2000, Cuernavaca, Mexico, 2000.

[8] Melamed, R. Green, and J. Turian, Precision and Recall of Machine Translation, In *Proceedings of the HLTNAACL 2003, Short Papers*: 61–63, Edmonton, Canada, 2003.

[9] Mishra, Vimal and Mishra, R. B., *Language Divergence between English to Sanskrit translation*, In Proceedings of COLIPS International Conference on Asian Language Processing (IALP), November 12-14, pp 175-181, Chiang Mai,Thailand, 2008.

[10] Mishra, Vimal and Mishra, R. B., Divergence handling in English to Sanskrit machine translation, International Journal of Information and Communication Technology , InderScience Publication, UK, 2009 (accepted).

[11] Papineni, Kishore & Roukos, Salim et al., BLEU: A Method for Automatic Evaluation of Machine Translation, Proceedings of the 20th Annual Meeting of the Association for Computational Linguistics, pp. 1– 9, 2002. [12] Satanjeev Banerjee and Alon Lavie, METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, In Proceedings of Workshop on “Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization” ACL 2005, pages 65-73, Ann Arbor, Michigan, 2005.

[12] Sinha, R.M.K. and Thakur, Anil, *Divergence Patterns in Machine Translation between Hindi and English,* 10th Machine Translation summit (MT Summit X), Phuket, Thailand, September 13-15, 2005.

[13] Sinha, R.M.K. and Thakur, Anil, *Translation Divergence in English-Hindi MT*, EAMT, May 30-31, Budapest, Hungary, 2005.

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

[15] Watanabe, H, Kurohashi, S. and Aramaki, E., Finding structural correspondences from bilingual parsed corpus for Corpus-Based Translation, Proceedings of COLING-2000, Saarbrucken,Germany,2000.