ORIGINAL RESEARCH

SAMRIDDHI Volume 12, Issue 1, 2020 Print ISSN: 2229-7111 Online ISSN: 2454–5767

Phrase-Based Statistical Machine Translation of Hindi Poetries into English by incorporating Word Sense Disambiguation

Rajesh Kumar Chakrawarti[1*], Pratosh Bansal[2], Jayshri Bansal[3]

[1]Ph.D. Scholar, Department of Computer Engineering, IET, DAVV, Indore, Madhya Pradesh, India

[2]Professor, Department of Information Technology, IET, DAVV, Indore, Madhya Pradesh, India

[3]Human Resource Development (HRD) Centre, DAVV, Indore, Madhya Pradesh, India

Abstract

Statistical machine translation (SMT) is a variant of machine translation where the translations are handled with statistically defined rules. Numerous researchers have attempted to build the framework which can comprehend the different dialects to translate from one source language to another target language. However, the focus on translation of poetry is less. Reliable and rapid transliteration of the poetry is very mandatory for the execution of the computer to translate the poem from one language to another. The existing approach has several issues, such as, time consumption, quality of the translation process, and matching of similar words. To overcome these issues, we propose a phrase-based statistical machine translation (PSMT) with special adherence to word sense disambiguation (WSD). The quality of the translation is increased by sensing the ambiguous words with WSD. The Hindi WordNet along with the Lesk algorithm identifies the ambiguous words and senses the exact meaning before the phrase extraction. Finally, the proposed method is compared with machine translation schemes, such as, rule-based machine translation and transfer-based machine translation. The experimental results suggest that the proposed method performed well with the inclusion of WSD in the PSMT technique.

Keywords: Hindi WordNet, Lesk algorithm, Machine translation, Phrase-based statistical machine translation, Word sense disambiguation.

SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology (2020); DOI: 10.18090/samriddhi.v12i01.11

Introduction

India being a multilingual nation, has various languages for communication. Yet Indians accentuate on Hindi for communication. Across the country, there is a large population that is not aware of linguistics and semantics of the English language. In this manner, there is a need to build up a machine translation application for the poems that will cross over any barrier between these two languages.[1,2] Today, most nations acknowledge Indian societies and attempt to get familiar with the Hindi language to learn the Indian culture. Hence, interpretations of Hindi writing and sonnets are extremely basic and significant to understand. The machine translation systems are automated computer programs (software) capable of translating information available in one language (called the source language) into different dialects (called the objective language).[3] With the support of machine translation frameworks, interpretations of Hindi writing into English are made simple. Existing frameworks on interpretations of Hindi sonnets into English is vital and an inconceivable exercise in machine translation model. In this, lyrics assume a significant role in contrast with the written work of interpretations. Since ballads give sentiments, feelings, expression, and more, the genuine interpretations of the sonnets are extremely significant.

The substantial data for the interpretation remained gathered to interpret the Hindi lyric into English. Since Hindi does not pursue any typical standard but, represented in various ways, a standard-based interpretation has been pursued, in which many linguistic principles have been built to actualize in a part of speech (POS) tagger. Moreover, Hindi words can be written in various ways but it

Corresponding Author: Rajesh Kumar Chakrawarti, PhD Scholar, Department of Computer Engineering, IET, DAVV, Indore, Madhya Pradesh, India, e-mail: rajesh_kr_chakra@yahoo.com

How to cite this article: Chakrawarti, R.K., Bansal, P. & Bansal, J. (2020). Phrase-based statistical machine translation of hindi poetries into english by incorporating word sense disambiguation. *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology, 12*(1), 54-61.

Source of support: Nil

Conflict of interest: None

does not pursue a particular spelling design. To address this problem, an accumulation of the corpus in an information base recognizes the right significance of phrase or word concerning its context by the WSD.[4,5] WSD has many applications, such as, document indexing, theme extraction, semantic annotation, genre identification, semantic web search, and information retrieval.[6-9] There are several techniques employed in machine translation. They are direct machine translation (DMT), rule-based machine translation (RMT), example-based machine translation (EMT), interlingua machine translation (IMT), knowledge-based machine translation (KMT), statistical-based machine translation (SMT), and hybrid-based machine translation (HMT).[10] Among various Machine Translation (MT) systems, the SMT plays a significant role in the translation process. SMT is a technique widely used for translation purposes with the help of statistical analysis in order to formulate rules which are best for the translation of a target sentence.[11] The SMT are of three different types, *viz*., word-, syntax-, phrase-, and hierarchical phrase-based. The word-based SMT utilizes the words and their neighbors during the translation process. While the phrase-based methods use phrases instead of words, it also considers the neighboring phrases while translation.[12] The syntax-based machine translation incorporates the syntax representation in order to find the best of the words. The hierarchical phrase-based is a hybrid approach that combines the strength of phrase and syntax-based methods. It employs the synchronous context-free grammar for translation purpose.[13,14] The proposed method employs the phrase-based SMT in order to solve the problems through the statistically devised rules. Some of the problems faced during the translation process are discussed as follows. The machine transliteration technique deals with the problem, such as, retrieval of cross-lingual, probability evaluation of translation, difficult to find the sentence with the prime and highest probability, and multiple representations of one word.[13] In the trans tech system, several words and grammatical rules arise the problem for translating into another language.[15] The example-based machine translation is good for translating the short sentence but it is worst for translating long sentences of the poem.[16] The word substitution in the hybrid approach does not provide the desired results as it does not care about the syntactic and semantic constraints of the target language.[17,18] The machine translation system needed to enhance the performance of the system with complex sentences. The major problem that arises in machine translation is due to the unavailability of structural, morphological differences, and word-aligned data during the translation of different languages.[19,20]

Contribution

In this paper, we aim to discuss the machine translation system based on the SMT. The study on various SMT and translation models of morphologically rich languages has been carried out. The study gives an insight into how the translation models have been carried out in each of the works. The research finds various challenges associated with the translation of Indian languages into English. One of the challenges is the words sense ambiguity widely found in the Indian language. The sense ambiguity occurs when certain words sound and spell similar but the sense of the word differs based on the context of the sentence. Generally in SMT, this problem is not addressed in most of the research. The proposed phrase SMT solves this problem by integrating the WSD with the SMT. The proposed strategy helps to improve the quality of translation.

Organization

The remaining section of the paper is described as follows:

- Section 1 discusses the background information.
- Section 2 provides the literature work in the machine translation system.
- Section 3 explains the proposed PSMT method in detail.
- In section 4, we present the result and analysis of the proposed Hindi to English translation model.
- Finally, section 5 concludes the work based on our result analysis.

Related Work

A few sorts of research have been carried out for the poetry translation, accessible in one language to other languages around the globe. Machine translation has rolled out a major improvement in making the Indian language progressively adaptable to learn and comprehend which has been considered by different translation techniques but has to face a few difficulties during translation.[21] The research on the SMT and the details of various translation models carried out are discussed in this section.

Xiong *et al*.[22] proposed a maximum entropy-based segmentation model for STM. The sentences are spitted into sequences of segments that can be translated. The phrasal extraction is a small module among the collection of modules used in the SMT. After extracting the phrases by maximum entropy the result is integrated with the SMT. The experiment is conducted with the news domain and is the method that has improved the quality of translation in terms of Bilingual Evaluation Understudy (BLEU). Ilknur and Kemal[23] proposed PSMT with the aid of local word reordering. The local word reordering concentrated to obtain the word order of certain English prepositional phrases and verbs with respect to the morpheme order of corresponding Turkish verbs and nouns. A morpheme-based language model is used for decoding and with the aid of word-based model re-ranking of n best list is handled by the decoder. The decoder output is repaired by correcting the words which have problems like incorrect morphological structure and words which are outside of the vocabulary. Liu *et al*.[24] proposed a framework where the translation memory is joined with the phrase-based statistical machine translation. The translation memory is integrated with the phrase-based MT through this approach. In the unified framework, different information is extracted from the translator machine (TM) in order to ease the SMT decoding process. The experiment is simulated in a Chinese

English TM database. The result shows that there is an improvement in BLEU score. The approach is tested with different models and training data in order to prove the sturdiness of the approach. Pathak *et al.*[25] proposed a method to translate from Hindi to English with an automatic parallel corpus generation system. In this mechanism, the Hindi news was translated into English news by google translator API and each of the translated English news headlines extracts its English content by the links provided from the google search which had its best token sort ratio from the weight of two phrases. Then align both the news and compare it with the fuzzy string matching algorithm by finding its similarity which was based on Levenshtein distance. Finally, the threshold value was taken from the matching algorithm, and if the ratio was more than the threshold value save both the news content as a part of the parallel corpus process. This approach needs to augment the parallel corpus for the translation among two languages. Sharma and Mittal[26] proposed a dictionary-based query translation system that aims to translate the Hindi words into English. In this technique, first, tokenize the queries, and before removing the stop words create the multi-word terms by using the n-gram technique for the translation. Next, match the source language terms with the bilingual dictionary if the terms are not translated then move to other cases by calculating the percentage match of the highest common subsequence of the source query. If the query terms are not translated in both the cases then, out of vocabulary (OOV) term was used by a rule-based approach would convert them into a roman format. Finally, collect the target language document from the input dataset and the query term was mapped to match the words for the translation. This approach was difficult to translate the named entities into the English language. Subalalitha *et al.*[27] proposed a template-based information extraction (IE) framework for translating the Tamil poem to English. In this framework, the information can be extracted by template-based information and n-grams based information. Initially, bilingual mapping was used to translate each word and two features are required to design templates, such as, term-based and Universal Networking Language Knowledge Base (UNL KB), which are used to match the words. In the term based features, there would be an appropriate match between the input and the feature. In some cases, some features were not present in the template then, UNL KB is used to add semantic constraints for the poem. Finally, the part-of-speech of the words presented in the poem is identified by the Tamil morphological analyzer that analyses the noun, verb, adverb, and adjective, which extracts the information by n-grams to check its grammar for the translation. This approach has high computational complexity for the translation process. Natu *et al.*[28] introduced a transfer-based machine translation (TBMT) scheme that aims to translate the Hindi text into English. This scheme performed several phases, such as, preprocessing, tokenizer, POS tagging, translation, and grammar check. Initially, the translated sentence should be sophisticated in the pre-processing stage and each word required POS tagging should be segmented into tokens. Then, the Viterbi algorithm used with the hidden Markov model assign sequence of POS tag for the translated words. The segmented words find their translated word from the database for translation. Finally, check the grammar for the words and arranged to determine the final structure of the translated words. This approach needs to enhance the quality of translation gender, number, and tense. Mishra *et al.*[29] presented the rule-based machine translation (RBMT) scheme to translate Hindi idioms into English. This technique contains two phases, such as, comparison and translation phases. In the first phase, the system compares the words with the comparison algorithm where the input was searched by the Hindi database. If the input was present in the database then input belongs to case 1, such as, different meanings and different forms in both the language and send to transfer-based module. Otherwise, it belongs to two cases, such as, case 2 and case 3, where input idioms with the same meaning and same form, others have the same meaning but different form in both the language in the database, in such a case send them to an interlingual-based module. The transfer-based module is composed of a tokenizer, parser, POS tagger, and declension tagger which translates the idiom by case 2 and 3. The interlingual-based module is composed of input, mapper, database, transfer-based module, and output which translate by case 1. Finally, these modules are used to produce the English translation. This approach needs to enhance the efficiency of the translation of idiom.

Proposed PSMT Method

The details of the proposed SMT is discussed in this section. Generally, SMT based translation methods consist of a collection of small modules that are involved in the translation process. The PSMT is used to translate the Hindi poetry along with incorporating the WSD approach to sense the disambiguation. The PSMT makes use of the phrases of one or more words in the translation process. The PSMT model first divides the input into phrases. The PSMT is based on the noisy channel model in the information theory.

Figure 1 shows the block diagram of the proposed approach. Consider a sentence *S* of a target language. The sentence *S* consist of a series of words. The machine translation transfers the source language which is given as $s_1^J = s_1,.....s_j.....s_J$ into a target language sentence $t_1^I = t_1,.....t_i.....T_I$. In SMT, the conditional probability is given as $P_r(t_1^I \mid s_1^J)$. The probability model is used for translation by finding a solution for the maximization problem.

$$t1^I = \arg\max_{t_1^I}\{P_r(t_1^I \mid s_1^J)\} \quad (1)$$

After applying Bayes theorem, equation 1 is given as

$$= \arg\max_{t_1^I}\{P_r(t_1^I)\, P_r(s_1^J \mid t_1^I)\} \quad (2)$$

Equation 2 provides the language model $P_r(t_1^I)$ and the target model $P_r(s_1^J \mid t_1^I)$. The phrase-based translation utilizes

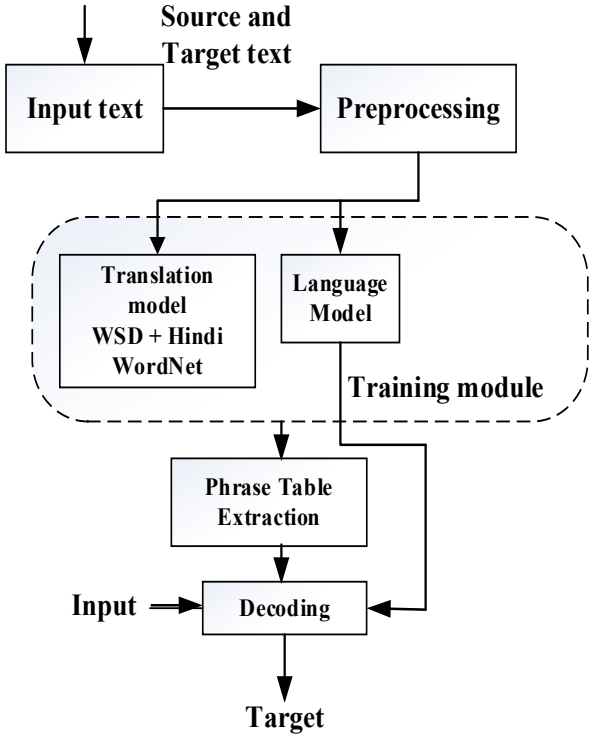

Figure 1: Proposed PSMT technique

the phrase which is a sequence of words. In PBT, the source sentence is segmented into phrase, and each phrase is translated and the target sentence is obtained from the phrase translation. The phrase translation probability is given as

$$P(\tilde{s} \mid \tilde{t}) = N(\tilde{s}, \tilde{t}) / N(\tilde{t}) \quad (3)$$

Where $N(\tilde{s}, \tilde{t})$ *and* $N(\tilde{t})$ denotes the count of event that translation has been done and count of phrase $\tilde{t}$. The $\bar{s}$ and $\bar{t}$ are considered to be subsequence in the source sentence *s* and target sentence *t* if both of them are longer than the k words.

Preprocessing

In the pre-processing phase, it focuses to solve the problems in the Hindi poem, such as, space inclusion (SI) and space deletion (SD) problem. Based on the context of a word, we expel such extra spaces from the content. Hindi is a Unicode support language. It implies that Hindi content cannot be handled utilizing ASCII coding. The raw data in our corpus had various character encoding schemes. Subsequently, before doing any further processing all the data was changed over to Unicode text file (UTF). The data gathered from the news sites were in the hypertext markup language (HTML) format so it was changed over to Unicode text files format. The portable document also changed over to text format. In the pre-processing stage, extra spaces and character encoding are resolved by using the AntConc tool from the annotated corpus. With the help of the tokenizer, the Hindi poem is divided into segments by sequence of words that require POS tagging into units known as tokens.

मछली जल की रानी है,
जीवन उसका पानी है |
हाथ लगाओ डर जाएगी,
बाहर निकालो मर जाएगी |

The segmented output (tokens): [' मछली ', ' जल ', ' की ', ' रानी ',' है ']

Word Alignment Model

The word alignment is the second step after preprocessing the words. In this proposed work, GIZA++ implementation of the IBM models is employed to perform the word alignment process. The tool runs the algorithm from source to target as well as, target to source in both of the directions. The IBM model evaluates the word to word probability of word to word alignment for all the source and target word for given sentences. The alignment is generated with a series of successive estimation in order to obtain a quality word alignment. The process needs several hours in order to process thee sentence with a higher number of quantity of sentences. The results of the alignment method create a link between the source and target words. The $\bar{s}$ is a source phrase, $\bar{t}$ is the target phrase, and $(\bar{a})$ is the alignment between source and target which is led by *a*, is valid only if the points do not cross the boundary of the bi-phrase.

Phrase Table Creation

Once the phrases are extracted with the GIZA++ word alignments, Moses directly uses the contents in order to generate a phrase table. The sequence of words is stored in phrase table as bi-phrases $(\bar{s}, \bar{t})$ along with the alignment $(\bar{a})$, if the following conditions are satisfied.

a. The alignment $(\bar{a})$ between the words is lead by "a" provided the condition that if there is at least one link and all links from a have both of their ends $(\bar{a})$ with or none.
b. The $\bar{s}$ and $\bar{t}$ are word subsequences in the source sentence x and target sentence y, respectively, and both should not have longer words than K number of words. The phrase extaction finds same words in both of the sentence. For example, consider the two sentences:

राजू के पास एक महंगा कलम ह, (rahu kae pass kalam hai) and नाई ने कलम को काटा (naee nae kalam lo kada). Both of these sentences have the same word kalam. Therefore, there arises a need to exactly sense the meaning of the word. Therefore, the WSD is integrated with the SMT. Once the words are disambiguated the phrases are extracted.

Integration of WSD with SMT

The sense of disambiguation is an important step in the translation process. There are many words that sound similar but their meaning varies based on the sentence. It is found that the word *kalam* has different meanings based on the meaning of the text. The meaning of *kalam* resembles pen, as well as, trimming based on the context of the sentence. Therefore, it is necessary to distinguish the meaning of the word before reordering the phrase-based SMT method. A knowledge-based method with the help of

Hindi Wordnet and Lesk algorithm is employed to sense the unambiguousness. The Hindi Wordnet is a database with a collection of nouns, adjectives, and adverbs. The following are the attributes of the Hindi WordNet are synset, gloss, and semantic relations.

- The synset is a collection with the same meaning. The contents of the synset are arranged based on usage popularity.
- Gloss consists of two parts, one where the texts are defined and the other which defines the importance of a word in a sentence.
- The semantic relation supports the relationship between the form and meaning of a word.

Lesk Algorithm

The Lesk algorithm was proposed by Michael E. Lesk in the year 1986. A simplified form of the Lesk algorithm is used in the proposed method. The sense of the words is decoded with the principle that the meaning of the word is decided by finding the sense which overlaps the most in WordNet and the given context. The structure of the Lesk algorithm explains how this algorithm works.

function LESK Algorithm (word, sentence) provides the exact sense of the word
exact sense<- words sense which occur frequently
best-sense <- continuous occurrence of a sense for word
maximum -overlap <-0
context <- words in a sentence
for each of the sense in a word senses do
signature <- collection of words in the gloss and sense examples
overlap <- CALCULATE OVERLAP (signature, context)
if overlap > maximum overlap then
maximum of the overlap <- overlap
best of the sense <- sense
end return (best of the sense)

The Lesk algorithm helps to find the correct meaning of words in a context through an individual decision by locating the sense which overlaps the maximum between the dictionary or WordNet definition. An example gives more ideas regarding the working of the Lesk algorithm. If the word *kalam* is searched on the WordNet, the senses of the word are found out. The sense which is maximum overlapped will be the output of the algorithm. Consider the example, राजू के पास एक महंगा कलम ह. It means that the person named Raju has a costly pen. The words are searched in the Hindi WordNet. In the sentence, the words like पास (*pass*), महंगा (*mahanga*), and कलम (*kalam*) have different senses based on the context. The search on word net shows various meanings for each of the words. The different senses for the words *pass*, *mahanga*, and *kalam* are provided in Figure 2.

The Lesk algorithm helps us to find the actual words based on the usage in a sentence. According to the Lesk algorithm, the best sense of *kalam* is pen. Similarly, other words are also processed and matched with the correct sense.

Once the word unambiguousness are identified, the phrase extraction is handled. The Figure 3 shows the extracted phrases of the poem.

मछली जल की रानी है,
जीवन उसका पानी है,

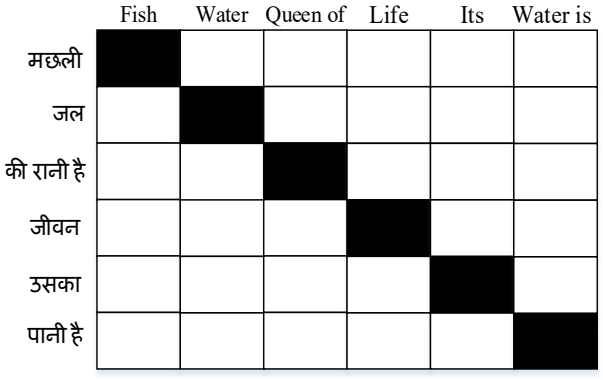

Figure 3: Phrase table

Figure 2: Sense for the words: **(a)** *Kalam*; **(b)** *Pass*; **(c)** *Mahanga*

The features of the phrase table generated with Moses are as follows. The Moses phrase table consists of five features for each bi-phrase. They are phrase translation probability, lexical weighting, phrase inverse translation probability, inverse lexical weighting, and phrase penalty. The value of the first four features takes values between 0 and 1. The decoder uses the features directly in order to generate a phrase table. With the aid of the tool Moses, the phrase pairs are extracted and the phrase pair score is calculated according to equation 3.

Re-ordering Model

The reordering technique is adopted from a Novel Reordering Model for Statistical Machine Translation developed in 2013.[30] The technique utilizes the phrasal dependency tree in order to order the translated words. The dependency relation between the contiguous synthetic non-syntactic phrases is used in the model.

Language Model

The language model employed in this proposed work is the n-gram language models, which are obtained from the connecting phrase-based continuous space language models (CSLM) growing method. The translation output can be considered as a concatenation of phrase in the phrase table. The connecting phrases are constructed based on the steps followed in Continuous space language models for statistical machine translation.[31]

The probability of a target phrase in SMT is given as

$$P_T(e) = \sum_f P_S(f) * P(e/f) \quad (4)$$

Where $P_S(f)$ is the probability of the source phrase. The connecting phrases are created at first and then employing the average probability of the connecting phrases is employed to take a decision such that which of the connecting phrases should be used.

For k gram connecting phrase, $w_1^n w_{n+1}^k$ where $w_1^n \in \beta w_1^n$ and $w_{n+1}^k \in w_{n+1}^k \gamma$. βw_1^n and $w_{n+1}^k \gamma$ are the phrases in the phrase table. The probability of the connecting phrases is given as

$$p_{connecting}(w_1^n w_{n+1}^k) = \sum_{n=1}^{k-1} (\sum_{\beta} P_t \arg e\ (\beta w_1^n) * \sum_{\gamma} P_t \arg e\ (w_{n+1}^k \gamma)) \quad (5)$$

The n-grams are generated based on steps like splitting, replacing, and renormalizing. An n-gram pruning method based on the phrase table is constructed based on the two conditions that the phrase table already consists of the contents and the contents are the result of the concatenation of two or more phrases in the phrase table. The connecting phrases are created based on the probability, as in equation 5. with the help of the threshold for each of the probability, which is higher than the threshold is retained.

Result and Analysis

The datasets for the proposed system consists of three text files where the parallel data set consists of two text files with 5,000 lines, where one is for Hindi poetries and the second is for translating into English poetries. The third one consists of monolingual data set containing around 10,000 lines for English poetries. The Moses tool is employed to evaluate the performance of the proposed system. Three performance metrics, *viz.*, precision, recall, and BLEU score are used in the proposed method. The precision method is a general evaluation method in MT. It is calculated from the number of correct words and the output length of the MT. The precision is given as

$$Precision = Number\ of\ correct\ words\ /\ Output\ -\ Length\ of\ Translation \quad (6)$$

The recall is obtained from the number of correct words divided by the reference length. The recall is given as

$$Recall = Number\ of\ correct\ words\ /\ Reference\ -\ Length\ of\ Translation \quad (7)$$

The BLEU score is used to find the accuracy of the proposed approach for the translation of Hindi poem into English poem and the evaluation is based on the geometric mean of the modified n-gram precision *p*, effective corpus length *r*, poem translation length l, and *N* is the number of words.

$$BLEU = \min\left(1 - \frac{r}{l}\right) + \sum_{n=1}^{N} W_n \log p_n \quad (8)$$

The implementation details are provided as follows. With the help of the python interface in Moses, the Lesk algorithm is implemented. We evaluate the MT output with the python based evaluation tool since the MT evaluation system does not provide an interface to evaluate the output. Thus, we calculate the precision and recall through the designed evaluation tool. Moses possesses a BLEU scoring tool named multi-bleu.perl. Moses also has another popular tool named NIST mteval script. The text needs to be converted into SGML format. The proposed system employs the NIST mteval script for calculating the BLEU score.

The proposed PSMT method is compared with TBMT and the RBMT translator, as shown in Table 1. The precision metric increased from 83.11 to 94.02%, and the recall metric improved from the range of 87.45 to 93.3%. The BLEU metric increases from 0.0695 to 80.04%. The proposed system provides more accurate results than the other systems.

The proposed PSMT's result is compared with RBMT and TBMT with the same data set used for the PSMT. In order to study the effect of WSD in the proposed method, the simulation is carried with and without WSD, as shown in Table 2.

Table 1: Performance metrices comparison

MT	*Precision*	*Recall*	*BLEU*
TBMT	83.11	87.45	0.0695
RBMT	92.05	91.09	0.7663
PSMT	94.02	93.3	0.9024

Table 2: Effect of WSD on the proposed method

Method	*Precision*	*Recall*	*BLEU*
Proposed with WSD	84.05	82.09	0.726
Proposed without WSD	83.02	93.3	0.9024

Table 3: Proposed approach *vs.* TBMT and RBMT for Hindi poetry translation

S. No.	*Hindi poetry*	*TBMT*	*RBMT*	*Proposed approach*
1.	चांद मामा आओ ना, ढूड-बटासा ना हो ना, मीठी लोरी गाओ ना, बिस्तार में है ना ना	Chand Mama Ao Naa Naa, Dhood-Batasa Nana, Sweet Lullaby Sing Naa, Bistar is not.	Come on, visit the moon, don't have to go to batsa, you're not in the sweet, but not in the bihar	The moon may not be able to come, be able to find a sweet, sweet, little gait, it is in the bed is n't it.
2.	बादल राजा, बादल राजा जलदी से पनि बरसा जा। नन्हे-मुन्ने झूलों में हैं। धरती की तू प्या भुज जा। जलदी से पनि बरसा जा	Cloud King, Cloud King Water may be showered with water. The little ones are in the swings. Go to the love arm of the earth. water-watering	The king, The King, pour over the clouds with the king of thunder. The little boy is in the swing. You go to Bhuj. Pani Pani Pani.	Rain king, rain the king come quickly to give water. The tiny ones are in dangles. Give you to the earth. rain down with water quickly
3.	हाथी आया, हाथी आया सूद हिलता है हाथ आया चलत तीर्थ हठ अया जूम जूम कर हाथ आया कान हिलता है हाथ आया	Elephant came, elephant came The trunk shakes the elephant came Walking Elephant Arrives Zoomed in, the elephant came Ear Shaking Elephant Arrives	The elephant came and the elephant came sniffing on the nose of the elephant, and the elephant came to him and the elephant came to him and the elephant came to him	The elephant came , elephant came into his hands, shaking his hands, and a mass of inlet-touched his hands, shaking his hands, shaking the hands, shaking his hands.

Table 2 shows the effect of WSD by comparing the proposed method with and without WSD. There is a change in the performance metrics when the WSD approach is applied. The proposed approach is examined with Hindi sonnet, which is translated into English interpretations in Table 3 on three machine translation framework.

Table 3 shows that the poems have been converted with good quality than the existing methods.

Conclusion

In the present era, machine translation is an important research in the natural language processing area. We have introduced a phrase-based STM for the interpretation of Hindi poem into the English language. We have developed a translator for Hindi to English language, which provides the best interpretation with the best precision. The approach has been successful in improving the quality of translation since the WSD model is also included to encounter the word sense ambiguity. The usage of the Hindi WordNet and Lesk algorithm has been able to provide the exact meaning of the words, as well as, improved the quality of translation. Future work integrates into a significant examination of the proposed component to deal with the complex Hindi/English sentence, different varieties of the same word, and consideration of increasingly linguistic sentences.

References

[1] Sinha, R. M. K. (2014, August). A system for identification of idioms in Hindi. In *2014 Seventh International Conference on Contemporary Computing (IC3)* (pp. 467-472). IEEE.

[2] Dutta-Roy, S. (2019). Negotiating Between Languages and Cultures: English Studies Today. In *English Studies in India* (pp. 61-72). Springer, Singapore.

[3] Alqudsi, A., Omar, N., & Shaker, K. (2014). Arabic machine translation: a survey. *Artificial Intelligence Review*, *42*(4), 549-572.

[4] Daud, A., Khan, W., & Che, D. (2017). Urdu language processing: a survey. *Artificial Intelligence Review*, *47*(3), 279-311.

[5] Bouhriz, N., Benabbou, F., & Lahmar, E. B. (2016). Word sense disambiguation approach for Arabic text. *International Journal of Advanced Computer Science and Applications*, *7*(4), 381-385.

[6] Sarmah, J., & Sarma, S. K. (2016). Decision tree based supervised word sense disambiguation for Assamese. *Int. J. Comput. Appl*, *141*(1), 42-48.

[7] Zhou, J., Yang, J., Song, H., Ahmed, S. H., Mehmood, A., & Lv, H. (2016). An online marking system conducive to learning. *Journal of Intelligent & Fuzzy Systems*, *31*(5), 2463-2471.

[8] Sreenivasan, D., Vidya, M., Sreenivasan, D., & Vidya, M. (2016). A walk through the approaches of word sense disambiguation. *Int. J. Innov. Res. Sci. Technol*, *2*(10), 218-223.

[9] Mittal, K., & Jain, A. (2015). WORD SENSE DISAMBIGUATION METHOD USING SEMANTIC SIMILARITY MEASURES AND OWA OPERATOR. *ICTACT Journal on Soft Computing*, *5*(2).

[10] Pathak, A. K., Acharya, P., & Balabantaray, R. C. (2019). A case study of Hindi–English example-based machine translation. In *Innovations in Soft Computing and Information Technology* (pp. 7-16). Springer, Singapore.

[11] Xiong, D., Zhang, M., & Li, H. (2011). A maximum-entropy segmentation model for statistical machine translation. *IEEE transactions on audio, speech, and language processing*, *19*(8), 2494-2505.

[12] Singh, M., Kumar, R., & Chana, I. (2019, August). Neural-Based Machine Translation System Outperforming Statistical Phrase-Based Machine Translation for Low-Resource Languages. In *2019 Twelfth International Conference on Contemporary Computing (IC3)* (pp. 1-7). IEEE.

[13] Singh, M., Kumar, R., & Chana, I. (2019, August). Neural-Based Machine Translation System Outperforming Statistical Phrase-Based Machine Translation for Low-Resource Languages. In *2019 Twelfth International*

Conference on Contemporary Computing (IC3) (pp. 1-7). IEEE.

[14] Mohaghegh, M., & Sarrafzadeh, A. (2012, March). A hierarchical phrase-based model for English-Persian statistical machine translation. In *2012 International Conference on Innovations in Information Technology (IIT)* (pp. 205-208). IEEE.

[15] Kaur, A., & Goyal, V. (2018, February). Punjabi to English Machine Transliteration for Proper Nouns. In *2018 3rd International Conference On Internet of Things: Smart Innovation and Usages (IoT-SIU)* (pp. 1-7). IEEE.

[16] Masroor, H., Saeed, M., Feroz, M., Ahsan, K., & Islam, K. (2019). Transtech: development of a novel translator for Roman Urdu to English. *Heliyon*, *5*(5), e01780.

[17] Ayu, M. A., & Mantoro, T. (2011, September). An Example-Based Machine Translation approach for Bahasa Indonesia to English: An experiment using MOSES. In *2011 IEEE Symposium on Industrial Electronics and Applications* (pp. 570-573). IEEE.

[18] Kaur, V., Sarao, A. K., & Singh, J. (2014). Hybrid approach for Hindi to English transliteration system for proper nouns. *International Journal of Computer Science and Information Technologies (IJCSIT)*, *5*(5), 6361-6366.

[19] Ye, Z., Jia, Z., Huang, J., & Yin, H. (2016, July). Part-of-speech tagging based on dictionary and statistical machine learning. In *2016 35th Chinese Control Conference (CCC)* (pp. 6993-6998). IEEE.

[20] Mall, S., & Jaiswal, U. C. (2013, September). Developing a system for machine translation from Hindi language to English language. In *2013 4th International Conference on Computer and Communication Technology (ICCCT)* (pp. 79-87). IEEE.

[21] Chakrawarti, R. K., Mishra, H., & Bansal, P. (2017). Review of machine translation techniques for idea of Hindi to English idiom translation. *International Journal of Computational Intelligence Research*, *13*(5), 1059-1071.

[22] Xiong, D., Zhang, M., & Li, H. (2011). A maximum-entropy segmentation model for statistical machine translation. *IEEE transactions on audio, speech, and language processing*, *19*(8), 2494-2505.

[23] El-Kahlout, I. D., & Oflazer, K. (2009). Exploiting morphology and local word reordering in English-to-Turkish phrase-based statistical machine translation. *IEEE transactions on audio, speech, and language processing*, *18*(6), 1313-1322.

[24] Liu, Y., Wang, K., Zong, C., & Su, K. Y. (2019). A unified framework and models for integrating translation memory into phrase-based statistical machine translation. *Computer Speech & Language*, *54*, 176-206.

[25] Pathak, A. K., Acharya, P., Kaur, D., & Balabantaray, R. C. (2018, September). Automatic Parallel Corpus Creation for Hindi-English News Translation Task. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1069-1075). IEEE.

[26] Sharma, V. K., & Mittal, N. (2018). Cross-lingual information retrieval: A dictionary-based query translation approach. In *Advances in computer and computational sciences* (pp. 611-618). Springer, Singapore.

[27] Subalalitha, C. N. (2019). Information extraction framework for Kurunthogai. *Sādhanā*, *44*(7), 156.

[28] Natu, I., Iyer, S., Kulkarni, A., Patil, K., & Patil, P. (2018, April). Text Translation from Hindi to English. In *International Conference on Advances in Computing and Data Sciences* (pp. 481-488). Springer, Singapore.

[29] Mishra, H., Chakrawarti, R. K., & Bansal, P. (2019). Implementation of Hindi to English Idiom Translation System. In *International Conference on Advanced Computing Networking and Informatics* (pp. 371-380). Springer, Singapore.

[30] Farzi, S., Faili, H., Khadivi, S., & Maleki, J. (2013). A Novel Reordering Model for Statistical Machine Translation. *Res. Comput. Sci.*, *65*, 51-64.

[31] Schwenk, H., Déchelotte, D., & Gauvain, J. L. (2006, July). Continuous space language models for statistical machine translation. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions* (pp. 723-730).