CHAPTER 5

HYBRID APPROACH FOR HINDI TO TAMIL TRANSLATION

5.1 INTRODUCTION

In this chapter, a hybrid approach is being proposed for the translation to further reduce loss of semantics during the pivot based Hindi to Tamil machine translation, besides that provided by word sense disambiguation. The ambiguity introduced by the use of pivot language has also been reduced by the use of semantic information during the translation process. Thus, the main aim of this hybrid approach is to handle the low resource availability issue without loss of the semantic information contained in the source text. The major phases designed in this hybrid approach besides preprocessing are,

i. Source language analysis
   a. Hindi part-of-speech tagger
   b. Word sense disambiguation

ii. Lexical transfer (source to pivot followed by pivot to target)

iii. Structural transfer

As used in the pivot-based approach, there is a lexical transfer phase which occurs in two steps – Hindi to English and then English to Tamil. But it was having semantic distortion due to which the overall performance is not up to the mark. The word sense disambiguation was introduced in the source language analysis phase, so that the sense information will be made available for the lexical transfer phase. The system architecture of this hybrid approach is shown in Figure 5.1. This system considers both the syntactic and semantic features to improve the accuracy of the machine translation system. These features are extracted during the source language analysis phase and are fed to the lexical transfer phase. The lexical transfer is performed on various identified relevant senses of the words mentioned in input sentence. The lexical transfer from Hindi to English is performed before the lexical transfer from English to Tamil. Once the lexical transfer is performed, the structural rearrangements is performed in the structural transfer phase. The structural
rearrangement is performed to keep the generated target sentence in a grammatically correct manner according to the target language grammar rules.

![Architecture of hybrid machine translation approach](image)

**Figure 5.1:** Architecture of hybrid machine translation approach

### 5.2 PRELIMINARY PHASE

The hybrid approach makes use of both pivot language and word sense disambiguation. The lexical transfer phase used in hybrid approach makes use of syntactic as well as semantic features. Thus, there is need for a translation model and language model which considers both the syntactic and semantic information. The translation model and language model has been developed for both language pairs – Hindi-English and English-Tamil. These translation model and language model are developed based on the parallel corpus provided by Technology Development for Indian Languages (TDIL) programme, Department of Information Technology (DIT), Ministry of Communication & IT, Government of India. The parallel corpus used has around 25000 sentence pairs and that too specifically on health domain. The lexical transfer phase can be mathematically represented as,
\[ P \left( \frac{w_t}{w_{s\text{pos}}} \right) = P \left( \frac{w_p}{w_{s\text{pos}}} \right) * P \left( \frac{w_t}{w_{p\text{pos}}} \right) \] 

(5.1)

Where,

\( w_s \) – Source language word

\( w_p \) – Pivot language word

\( w_t \) – Target language word

\( pos \) – Part-of-speech of source word

Which is same as in pivot-based approach except that the \( w_s \) mentioned in above equation is over the various identified relevant senses of the word occurring in the input source text. The relevant senses are identified using the word sense disambiguation. The target word whose probability is maximum will be the most appropriate target word for the given source word.

### 5.3 PREPROCESSING PHASE

The input sentence is subjected to tokenization at word level using a tokenization process. Once the words are tokenized, there is need for analysis on the words to gather the information about the morphology of the words. As in earlier approaches, the morphology of the words is extracted using longest affix matching algorithm which in turn uses the Levenshtein distance to find the nearest affix that can be stripped from the input word. The affixes which are separated from the root word should not be discarded since, they have the tense, aspect and modality (TAM) information. A list of affixes along with its information are maintained for identifying the affixes in the words. Once the morphological information of the words is extracted, the syntactic and semantic relation between the words need to be analyzed. The relation between words are analyzed and its syntactic feature is extracted in source language analysis phase.
5.4 SOURCE LANGUAGE ANALYSIS

This language analysis phase is divided into two sub-modules – part-of-speech tagger module and word sense disambiguation module. The Hindi part-of-speech tagger is used to extract the syntactic information from the given text. The word sense disambiguation phase is used to extract the semantic information from the source text. The performance of proposed hybrid approach has dependency over both the extracted information.

5.4.1 HINDI PART-OF-SPEECH TAGGER

Hindi part-of-speech tagger extracts syntactic features from the source text and it is developed using a multilayer perceptron-based neural network [62]. Since, this machine translation system basically uses syntactic and semantic information, the accuracy of the tagger is more important for the accuracy of translation. Hence, the multilayer perceptron-based part-of-speech tagger was used instead of simple hidden Markov model (HMM) based part-of-speech tagger. The input and output layer size of multilayer perceptron-based tagger will be same as the number of part-of-speech tags being used in the system. Only one hidden layer is used in this multilayer perceptron-based tagger. It has been found that with more number of hidden layers the performance degrades instead of improving. The hidden layer is activated using a sigmoid function [63] which is mathematically represented as below,

\[
y(f(x)) = \frac{1}{1+e^{-f(x)}}
\]  

\[
f(x) = \sum_{j=1}^{N} w_j * P(\text{word}_{tag_j}) * P(\text{tag}_j | \text{tag}_i)
\]

where,

- *word* – current word being processed
- *tag_i* – Part-of-speech tag of previous word
- *tag_j* – Part-of-speech tag from the considered POS tag set
- *N* – number of distinct part-of-speech tags
Both probabilities mentioned in \( f(x) \) are generated using the input corpus and are further fed to the multilayer perceptron-based tagger for training purposes. The various possible part-of-speech of the input word are predicted using the word’s position and its preceding word’s part-of-speech. Once it has been trained, the identified probable part-of-speech of a word is fed to the system as input and the system will provide a vector as an output. This vector is mapped with the part-of-speech to decode the exact part-of-speech of the given word.

### 5.4.2 WORD SENSE DISAMBIGUATION

Once the extraction of syntactic information is performed, the sentence is subjected to sense disambiguation phase to identify appropriate sense for the word mentioned in that context. This proposed system makes use of latent semantic analysis (LSA) to perform sense disambiguation [58]. The singular value decomposition is performed in a similar manner as mentioned in chapter-3 before. The term-frequency matrix is generated using sentences that uses the same input word but in different context. The various contextual sentences are retrieved from the Hindi wordnet [55]. The generated term-frequency matrix is decomposed to three different matrices and the similarity between sentences are calculated to identify the appropriate sense for the input word. To identify the similar sentence, cosine similarity is applied over the dot product of the right singular matrix and the singular diagonal matrix. The sentence which has maximum cosine similarity value is considered as the appropriate sense of input word mentioned. The identified syntactic and semantic information are used by the lexical transfer phase for translation purpose.

### 5.5 LEXICAL TRANSFER

The lexical transfer phase is used to translate the source word to its appropriate target word using the statistical information generated from the parallel bilingual corpus. Since there is an intermediate pivot language, the lexical transfer phase has to work in two steps – lexical transfer phase-I and lexical transfer phase-II. These two modules work in the same manner as mentioned in chapter-4. In addition to that, the lexical transfer phase makes use of the word sense information to reduce the number of possible ambiguous output which in turn reduces the semantic distortion that occurred in the pivot-based approach (without word sense disambiguation). The phase is mathematically expressed as,
Using extended Bayes theorem, the above expression (5.4) is rewritten as,

$$P\left( \frac{w_t}{w_s \cdot \text{pos}} \right) = \left[ P\left( \frac{w_s}{w_p} \right) \right] * \left[ P\left( \frac{pos}{w_p} \right) * P\left( \frac{w_p}{w_{p-1}} \right) \right]$$

(5.5)

The hybrid approach for Hindi to Tamil machine translation is compared with the naïve Bayes statistical machine translation system in terms of the features that are being used in both the system. The comparison shown in Table 5.1 illustrates about the advantages and disadvantages in both the system. Both the proposed hybrid approach and naïve Bayes statistical approach makes use of the language model and translation model. But, in hybrid approach, a pivot language is being used to handle the low resource availability issue, due to which the semantics may get lost while translating from source to target through pivot language. In order to handle this distortion, the word sense feature was introduced in the proposed approach which does not exist in the naïve Bayes approach. Due to the word’s morphological structure in Hindi and Tamil language, the part-of-speech of the word can be used as a feature to predict the appropriate translation. Thus, part-of-speech was also used in the proposed approach. The alignment model that was used in naïve Bayes approach doesn’t consider the part-of-speech of the words which was introduced in the proposed method since there is free word order nature in the languages, Hindi and Tamil.

Table 5.1: Comparison of proposed machine translation with Naïve Bayes statistical machine translation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Language Model $P(s)$</td>
<td>Used for source language</td>
<td>It considers preceding word during translation mechanism which improves the translation</td>
<td>Used for source language as well as pivot language</td>
<td>It considers preceding word during translation and pivot language is used as there is low resource availability</td>
</tr>
<tr>
<td></td>
<td>Translation Model ( P(s</td>
<td>t) )</td>
<td>Used for target language</td>
<td>It maps the source word with the target language and predicts the probable translation which also helps during translation</td>
<td>Used for target as well as pivot language</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>3</td>
<td>Part-of-speech (POS)</td>
<td>Not used</td>
<td>It is a disadvantage</td>
<td>Used in the translation model which is modified accordingly</td>
<td>POS provides more accurate mapping between source and target words. Thus, it improves the translation accuracy</td>
</tr>
<tr>
<td>4</td>
<td>Alignment Model</td>
<td>Used for word alignment</td>
<td>Uses traditional IBM models and doesn’t consider POS during alignment</td>
<td>Used for word alignment but has been modified to consider part-of-speech during alignment process</td>
<td>Uses modified IBM model and its alignment improves since the word position depends on the POS of the word as well</td>
</tr>
<tr>
<td>5</td>
<td>Word Sense</td>
<td>Since it works only on probabilistic manner, it does not include sense identification.</td>
<td>Since contextual information is not considered translation is poor in certain cases</td>
<td>During translation phase, it considers the words sense in the context it is being used</td>
<td>Improves the translation based on contextual information. Words translation differs based on the sense it is being used</td>
</tr>
</tbody>
</table>

### 5.6 STRUCTURAL TRANSFER

Since the natural languages being used are free word ordered, the hybrid approach requires the structural transfer phase. The structural transfer phase uses Naïve Bayes approach to identify the probable grammatical sequence for the target sentence generated by the lexical transfer phase. The features being used in this approach are – sequence of target words and part-of-speech of source word. Before reaching this phase, the syntactic and semantic information are extracted from the text. But based on language analysis, it is found that the syntactic information contributes to the position of the word in the target language whereas, the semantic information does not have any influence on this. Thus, only the syntactic feature is used in the rearrangement phase. The inclusion of syntactic feature in this rearrangement phase has reduced the ambiguity in the grammatical sequence. The maximum probable grammatical sequence is predicted using the alignment table constructed using GIZA++. 
5.7 RESULT ANALYSIS

For illustration, consider the source text as - एक सामान्य व्यक्ति का ब्लड प्रेशर 140 - 90 से कम होना चाहिए। (ek saamaany vyakti ka blad preshar 140 - 90 se kam hona chaahie.). The pivot-based approach discussed in previous chapter was applied on this source text and it had slight error in the output text. It is noted that this error can be rectified using the word’s sense. Initially, this source text is subjected to word level analysis using morphological analyzer. Since, the input text has only root words and doesn’t have any TAM information, there is no change in the input text after morphological analysis. This text is further fed to part-of-speech tagger and the output of the tagger is - एक/QT_QTC सामान्य/JJ व्यक्ति/N_NN का/PSP ब्लड/N_NN प्रेशर/N_NN 140/QT_QTC -/RD_SYM 90/QT_QTC से/PSP कम/QT_QTF होना/V_VAUX चाहिए/V_VAUX ।/RD_PUNC.

The word sense will be disambiguated with the help of latent semantic analysis (LSA). Consider the word “सामान्य (saamaany)” and its identified part-of-speech i.e., adjective. The word and its part-of-speech are used to extract the various possible senses from the Hindi wordnet. The various senses retrieved for the word “सामान्य (saamaany)” from the wordnet are – साधारण (sadharan), नामल (normal), सामूहिक (samuhik), सार्वनिक (saarvthrik), सादा (saada), आम (aam). All the sentences for these senses are retrieved from the wordnet. Retrieved sentences are mentioned below in the order of identified senses,

S1: यह सामान्य साड़ी है।

(yah saamaany sari hai.)

S2: शहर की हालत सामान्य हो रही है।

(shahar kee haalat saamaany ho rahee hai.)

S3: सामूहिक सभा का आयोजन किया गया।

(saamoohik sabha ka aayojan kiya gaya.)
The term-document frequency matrix is constructed using all the above sentences \([S1, S2, \ldots, S6]\) along with the input sentence. Using these 7 sentences, the number of distinct words is identified as 39. Thus, the term-document frequency matrix \((A)\) will be of size \((39 \times 7)\) and is as shown below.

\[
A = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix} \quad (5.6)
\]

Using singular value decomposition \([66]\), the term-document frequency matrix is decomposed into three different matrices – left singular matrix \((L)\), right singular matrix \((R)\) and singular diagonal matrix \((S)\). The left singular matrix will be of size \((39 \times 39)\), whereas the right singular matrix will be of size \((7 \times 7)\). The singular diagonal matrix will also be of size \((7 \times 7)\), but, it has non-zero values as diagonal elements. These three matrices are as shown below,

\[
L = \begin{bmatrix}
-0.2244 & -0.1560 & 0.0548 & -0.0176 & \ldots & -0.1078 \\
-0.4330 & 0.2353 & 0.0790 & 0.3091 & \ldots & 0.0460 \\
-0.2244 & -0.1560 & 0.0548 & -0.0176 & \ldots & -0.0196 \\
-0.2515 & -0.1931 & -0.3467 & 0.0426 & \ldots & -0.0071 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
-0.0560 & 0.0830 & -0.0137 & -0.2584 & \ldots & 0.8237
\end{bmatrix} \quad (5.7)
\]
The cosine similarity between vectors in right singular matrix and first row of singular diagonal matrix is calculated. The resultant vector after applying cosine similarity is as below,

\[
\text{Cosine similarity} = [0.13 \ 0.3341 \ 0.5423 \ 0.4866 \ 0.5335 \ 0.4971 \ 0.5086] \quad (5.10)
\]

The sentence which has the cosine similarity nearer to 0 is closest to the input sentence. The first value in the vector denotes the cosine similarity with the input sentence itself and it is natural that it will be closest to zero. The next smallest value in the vector is 0.3341 which is the cosine similarity value of sentence S1. The word’s sense used in S1 is साधारण (saadhaaran) and it is the most matching sense for the word सामान्य (saamaany) according to the context in which it is used.

**Table 5.2:** Transfer table for Hindi word “saadhaaran”

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Hindi Word</th>
<th>Part-of-speech Tag</th>
<th>English Word</th>
<th>Frequency of occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>साधारण</td>
<td>JJ (Adjective)</td>
<td>Simple</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>साधारण</td>
<td>JJ (Adjective)</td>
<td>Normal</td>
<td>12</td>
</tr>
</tbody>
</table>

The lexical transfer phase uses the identified semantically similar word साधारण (saadhaaran) instead of the actual word सामान्य (saamaany). The transfer table for the Hindi word “साधारण (saadhaaran)” is as shown in Table 5.2.
For instance, consider the English word (pivot) as “simple” whose frequency of occurrence in the corpus is 36. Out of these 36 occurrences, it was tagged as adjective in 24 instances. Consider the probable preceding word as “A”, which occurred 54 times in the corpus. The word “A” precedes word “simple” in 6 occurrences. Thus, the probable pivot word is found using below expression,

\[
P\left(\frac{w_p}{w_s pos}\right) = P\left(\frac{w_s}{w_p}\right) * P\left(\frac{pos}{w_p}\right) * P\left(\frac{w_p}{w_p-1}\right) = \frac{4}{36} * \frac{24}{36} * \frac{6}{54} = 0.0082
\]  

(5.11)

Considering the pivot word (in English) as “normal” whose frequency of occurrence in the corpus is 24. Word “normal” is tagged as adjective in 12 of its occurrences in the corpus. The word “A” precedes word “normal” in 2 occurrences. The probability of word “normal” as pivot word is calculated below,

\[
P\left(\frac{w_p}{w_s pos}\right) = P\left(\frac{w_s}{w_p}\right) * P\left(\frac{pos}{w_p}\right) * P\left(\frac{w_p}{w_p-1}\right) = \frac{12}{24} * \frac{12}{24} * \frac{2}{54} = 0.0092
\]  

(5.12)

Table 5.3: Transfer table for English word “normal”

<table>
<thead>
<tr>
<th>S. No.</th>
<th>English Word</th>
<th>Part-of-speech Tag</th>
<th>Tamil Word (Cātāraṇa)</th>
<th>Frequency of occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Normal</td>
<td>JJ (Adjective)</td>
<td>சாதாரண       (Cātāraṇa)</td>
<td>85</td>
</tr>
<tr>
<td>2</td>
<td>Normal</td>
<td>JJ (Adjective)</td>
<td>இயல்பாக (Iyalpāka)</td>
<td>10</td>
</tr>
</tbody>
</table>

From the above calculations, the most probable pivot language word for “साधारण/JJ (saadhaaran)” is “normal”. All the other words are also translated to their probable pivot language word and the resultant pivot language sentence is - “a/DT normal/JJ man/N_NN blood/N_NN pressure/N_NN 140/QT_QTC −/RD_SYM 90/QT_QTC less/QT_QTF than/PSP should/V_VAUX be/V_VAUX”. This intermediate text is fed to transfer phase-II and the calculation for translating the word from “normal” to its equivalent Tamil word is as follows,

The word “normal” can be mapped with the word “சாதாரண (Cātāraṇa)” and also with the word “இயல்பாக (Iyalpāka)”. Firstly, consider the probability calculation of word “சாதாரண (Cātāraṇa)” that occurs 116 times in the Tamil corpus. Out of these 116 occurrences,
the word “साधारण (Cātāraṇa)” occurs as adjective in 100 instances. Assume, the probable preceding word is found as “ஒரு (oru)” and it occurred 23 times in the corpus. The word “ஒரு (oru)” precedes word “சாதாரண (Cātāraṇa)” in 14 instances. The probability of target word “சாதாரண (Cātāraṇa)” is calculated as below,

\[
P \left( \frac{w_t}{w_p, p_{os}} \right) = P \left( \frac{w_p}{w_t} \right) * P \left( \frac{p_{os}}{w_t} \right) * P \left( \frac{w_t}{w_{t-1}} \right) = \frac{85}{116} * \frac{100}{116} * \frac{14}{23} = 0.384
\]  

(5.13)

The word “இயல்பாக (Iyalpāka)” occurred 10 times in the corpus and it is tagged as adjective in 5 of its occurrence. In none of its occurrence, it is preceded by the word “ஒரு (oru)”. Thus, the probability calculation for the Tamil word “இயல்பாக (Iyalpāka)” with the given pivot word “normal” will be,

\[
P \left( \frac{w_t}{w_p, p_{os}} \right) = P \left( \frac{w_p}{w_t} \right) * P \left( \frac{p_{os}}{w_t} \right) * P \left( \frac{w_t}{w_{t-1}} \right) = \frac{10}{10} * \frac{5}{10} * \frac{0}{23} = 0
\]  

(5.14)

Table 5.4: Analysis of generated target text with its reference text

<table>
<thead>
<tr>
<th>Generated Text</th>
<th>ஒரு (oru)</th>
<th>சாதாரண (cataran)</th>
<th>நபரின் (naparin)</th>
<th>இரத்த (iratta)</th>
<th>அழுத்தம் (aluttam)</th>
<th>140 – 90</th>
<th>குளறக்க (kuraikka)</th>
<th>வவண் டும் (ventum)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference Text</td>
<td>ஒரு (oru)</td>
<td>சாதாரண (cataran)</td>
<td>நபரின் (naparin)</td>
<td>இரத்த (iratta)</td>
<td>அழுத்தம் (aluttam)</td>
<td>140 – 90</td>
<td>குளறக்க (kuraikka)</td>
<td>வவண் டும் (ventum)</td>
</tr>
<tr>
<td>Match</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Based on the above calculation in expression (5.8) and (5.9), it is found that the word “சாதாரண (Cātāraṇa)” will be most probable translation for the word “सामान्य (saamaany). The structural transfer phase has identified that the sequence does not need any rearrangements to retain the target language grammar. The final translated text for this input sentence is shown in the Table-5.4 and it is also compared with the reference sentence mentioned in the corpus.
Table 5.5: Comparison of proposed hybrid approach with the proposed word sense-based approach

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Corpus Size (in number of words)</th>
<th>Word Sense Based Hindi-English-Tamil Machine Translation</th>
<th>Word Sense based Hindi-Tamil Machine Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Precision (in %)</td>
<td>Recall (in %)</td>
</tr>
<tr>
<td>1</td>
<td>10000</td>
<td>68</td>
<td>58</td>
</tr>
<tr>
<td>2</td>
<td>20000</td>
<td>70</td>
<td>69</td>
</tr>
<tr>
<td>3</td>
<td>30000</td>
<td>81</td>
<td>76</td>
</tr>
</tbody>
</table>

From the Table-5.4, the target sentence is found to be same as it has to be with respect to the reference sentence. The proposed hybrid machine translation system has been evaluated using various sentences (restricted to health domain) and the Bilingual Evaluation Understudy (BLEU) score is calculated to be 0.68. Precision and recall gradually increases with respect to corpus size as shown in Table-5.5. The corpus size when increased, further leads to increase in distortion noise and thus producing a poor translation accuracy. It has been found ideally that the unique word should be kept at 30000 and further increase in the corpus size leads to poor translation due to increase in noise.

The proposed hybrid approach has been compared with the word sense-based Hindi to Tamil machine translation system and also with the pivot-based Hindi to Tamil machine translation system. Figures 5.2 and 5.3 show the comparison of these systems in terms of precision and recall. It is very clear from Figure 5.2 that the precision of proposed hybrid machine translation system improves with respect to corpus size, but its precision is lesser when compared with word sense-based Hindi-Tamil MT system which is due to the increase in distortion by the inclusion of pivot language during translation. From Figure-5.3, it is visible that the recall is relatively good compared to the pivot-based Hindi-Tamil MT. But it degrades when compared to the word sense-
based Hindi-Tamil MT due to the distortion caused by the introduction of pivot language in between the source and target language translation.

**Figure 5.2:** Comparison of Hindi to Tamil Machine Translation in terms of precision

The proposed system is also compared with the other statistical machine translation system in terms of BLEU (Bilingual Evaluation Understudy) score, which is shown in Table-5.6. The hybrid approach is found to have a BLEU score of 0.7637 and it is also noted that the BLEU score has
improved by few percentage when compared with the pivot-based approach discussed in chapter-4.

Table 5.6: Comparison of various statistical machine translation system using BLEU score

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Methodology</th>
<th>Source language</th>
<th>Target language</th>
<th>BLEU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Statistical machine translation [67]</td>
<td>English</td>
<td>Bahasa Indonesia</td>
<td>0.2287</td>
</tr>
<tr>
<td>2</td>
<td>Lemma translation [68]</td>
<td>Japanese</td>
<td>Indonesian</td>
<td>0.1282</td>
</tr>
<tr>
<td>3</td>
<td>Lemma translation [68]</td>
<td>Indonesian</td>
<td>Japanese</td>
<td>0.1723</td>
</tr>
<tr>
<td>4</td>
<td>Proposed pivot-based approach</td>
<td>Hindi</td>
<td>Tamil</td>
<td>0.7394</td>
</tr>
<tr>
<td>5</td>
<td>Proposed hybrid approach</td>
<td>Hindi</td>
<td>Tamil</td>
<td>0.7637</td>
</tr>
</tbody>
</table>

5.8 CONCLUDING REMARKS

It is evident from the results of the hybrid system that the noise introduced in a pivot-based machine translation system has been reduced by the use of word sense disambiguation along with the pivot-based system. But as compared with the word sense-based Hindi to Tamil machine translation system, the performance of hybrid system is still poor. This is due to the deviation that is happening by the use of pivot language and the language specific properties. Hindi and Tamil language are morphologically rich language when compared with English language. Thus, the morphological structure of words differs between source language, target language and pivot language. This difference leads to loss of semantics during translation and therefore, produces a poor translation than the one generated by direct statistical machine translation (without pivot). These bottlenecks can be addressed by the use of deep learning architecture to capture the semantics and syntactic features of the languages to perform machine translation. The deep learning approach is the topic of the next chapter.