CHAPTER 7

CONCLUSIONS AND FUTURE WORK

7.1 CONCLUSIONS

In today’s multicultural business world, there is involvement of more natural languages for establishing communication in the business environment. Due to globalization, there is need for machine translation system to assist in communication between two different organization. Thus, the demand for translation system on various language pairs has increased. It can also help in improving the communication between people from different origin. The development in machine translation also promotes the re-establishment of various business activities according to the national resources available for the business. For example, a Canadian based MNC can establish its office in India if they get proper translation services. To achieve this, there is prime need for developing machine translation system with more accuracy. In the current scenario, the machine translation system between Indian languages are still in need of development due to its poor accuracy. Similarly, there is need for machine translation services between prominent languages of northern and southern India for more business/technical interactions.

Developing a machine translation system between Indian language pair is the primary and major issue that needs to be addressed. The Indian languages such as Hindi and Tamil have poor resource availability. Existing translation system on this language pair has focus towards syntactic features of the languages. But, there is need to consider the semantic features of the languages too. Thus, a combination of both syntactic and semantic feature will provide a more accurate machine translation.

The next major challenge was to generate the parallel corpus that is required for a machine translation system. Since, the availability of resources in Hindi and Tamil is very poor, there was need for some intermediate pivot language to assist in the translation. This research work reports about one such pivot-based approach which uses English as the pivot language due to its vast resource availability.
One more major challenge in the research was to improve the overall accuracy by making use of the language specific features. Hindi is morphologically rich and partially free word ordered language. Whereas, Tamil is morphologically rich but fully free word ordered language. Since both the languages are morphologically rich, the tense, aspect and modality information are stored along with the root word. During translation, this information also plays major role. Apart from this, the word order also contributes to an accurate translation. All these language specific features have to be extracted and used for improving the performance of a translation engine.

The target sentence being generated should follow the grammar rules of target language. Grammar followed by Hindi and Tamil languages are in the subject-object-verb (SOV) form. The Hindi sentences are partially free word order whereas, Tamil sentences are fully free word order. This word order feature increases the challenge in mapping the target text according to its grammar rules.

The following are list of observations and contributions in this research on Hindi to Tamil machine translation,

1. Word sense-based approach for Hindi to Tamil machine translation was proposed, which considers both the syntactic and semantic feature of both languages. Syntactic features contribute to the mapping of sentences with its probable target text. But the semantic feature provides more detailed information about the most appropriate target word for a given source word. It is observed that the approach has issue with the syntactic information being retrieved. This is due to poor performance by the HMM based part-of-speech tagger. Even the amount of resources needs to be increased for better performance of this approach. But, both Hindi and Tamil are poor resource languages.

2. A pivot-based approach for translation was proposed to handle the low resource issue. Since, English has vast resources as compared with Hindi and Tamil, English is used as a pivot language. To improve the performance of syntactic process, a multilayer perceptron based neural network was used to perform part-of-speech tagging. The multilayer perceptron based neural network extracts the syntactic information from the
input sentence and this information is used during the translation process. But Hindi and Tamil are morphologically rich languages as compared with the morphology of English. Due to which there is loss of semantics and it is being handled by the introduction of sense disambiguation phase in a pivot-based approach.

3. A hybrid approach was proposed to handle the semantic distortion that occurred in a pivot-based approach. This approach uses word sense disambiguation and the identified senses aid in the translation process. The syntactic information extracted using multilayer perceptron tagger is also used in this hybrid approach. The improvement in performance of Hindi to Tamil translation by the use of hybrid system saturates after particular threshold on corpus size. Due to difference in word morphology in Hindi, English and Tamil, there is introduction of semantic distortion during translation. Thus, the actual information gets lost during translation.

4. In order to improve the translation accuracy further, a deep learning approach was proposed. This approach learns the features from the parallel text fed to it. The parallel text was embedded into a vector without losing the syntactic and semantic features of the languages. Deep learning approach uses attention mechanism to extract the features that can help in sentence rearrangement process.

The machine translation approaches were evaluated using Bilingual evaluation understudy (BLEU). The BLEU score of proposed approaches have been calculated and are listed in the Table 7.1.

From the table, it is visible that the performance of all the approaches except pivot-based approach is better. This is due to semantic distortion that occurred by the usage of three different languages. The performance of hybrid approach is better when compared with all other approaches. But, its performance does not improve after a particular threshold value even with increase in the corpus size. Out of all the approaches, the deep learning approach has performed better when compared with word sense-based approach and pivot-based approach, and it can be improved further with increase in corpus size.
Table 7.1: Comparison of various proposed machine translation approaches in terms of BLEU score

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Approach</th>
<th>BLEU score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Word sense-based approach</td>
<td>0.6800</td>
</tr>
<tr>
<td>2</td>
<td>Pivot-based approach</td>
<td>0.5400</td>
</tr>
<tr>
<td>3</td>
<td>Hybrid approach</td>
<td>0.7637</td>
</tr>
<tr>
<td>4</td>
<td>Deep learning approach</td>
<td>0.7588</td>
</tr>
</tbody>
</table>

7.2 FUTURE WORK

The significant and prominent improvements that can further increase the accuracy of Hindi to Tamil machine translation system are listed below:

1. The word sense-based Hindi to Tamil statistical machine translation system can further be improved by the introduction of more accurate part-of-speech tagger and word sense disambiguation. The increase in corpus for this approach may also improve the overall accuracy of the system.

2. A multilayer perceptron-based part-of-speech tagger was designed using the statistical features. The performance of tagger with one hidden layer is found to be better as compared with the network having more than one hidden layer. The statistical feature being used for training the network is based on a bigram language model.

3. Pivot based statistical machine translation can also be improved by the introduction of more accurate part-of-speech tagger. The performance of this approach may be improved by using a pivot language which has high resource availability and has some relation with either the source or the target language. Thus, the use of a different pivot language may be explored.

4. The hybrid approach uses the word sense module along with the pivot-based approach. This approach can also be improved by using a pivot language which has relationship with either the source or the target language. Even the improvement in accuracy of word sense
disambiguation phase can provide a considerable improvement on the machine translation system.

5. The neural machine translation system proposed in this research can have further improvements by the introduction of multi-sense-based word embedding instead of single-sense word embedding model. The sequence to sequence model can also be improved by using dual encoder and decoder in it, which may help in improving the learning by the network.

6. The sequence to sequence machine translation system can also be modified by the introduction of pivot language in it. This may help to improve the overall accuracy even with lesser resources.

On the whole, the future improvements on Hindi to Tamil machine translation system can be made by increasing the parallel corpus and by applying the transfer-based approach along with some other machine translation approaches. These approaches can also be extended further on the phrase level to maintain the relationship between words. The alignment phase which was used in the statistical approaches can be modified such that it can consider the global association instead of the local association of word in the sentence.