Deep Learning Approach to English-Tamil and Hindi-Tamil Verb Phrase Translations

D. Thenmozhi, B. Senthil Kumar and Chandrabose Aravindan

Department of CSE, SSN College of Engineering, Chennai
{theni_d,senthil,aravindanc}@ssn.edu.in

Abstract. Verb phrase (VP) translation focuses on translating all forms of verbs that helps in Machine translation (MT) task. This has several applications such as cross lingual information retrieval (CLIR), speech synthesis, natural language understanding and generation. VP translation is a challenging task due to variations of characteristics, structure and families among the languages. Further, developing a language independent methodology for VP translation is an interesting task. In this paper, we present a deep learning methodology for English-Tamil and Hindi-Tamil VP translations. We have adopted neural machine translation model to implement our methodology for VP translation. Our approach was evaluated using the data set given by VPT-IL@FIRE2018 shared task.

Keywords: Verb Phrase Translation · Machine Translation · Text mining · Deep Learning · Indian Languages · Tamil Language.

1 Introduction

Verb phrase (VP) translation is part of Machine translation (MT) task which focuses on translating all forms of verbs such as main verb, auxiliary verb, finite verb, non-finite verb and negation verb. This has several applications such as MT [10, 3], cross lingual information retrieval (CLIR) [12, 13], speech synthesis, sentence simplification [5], natural language understanding and generation. VPs carry several information like tense, modal and person-number-gender (PNG). VP translation is a challenging task due to the characteristics that vary from language to language. Some languages such as Tamil, Hindi and Telugu have subject-verb agreement and other languages such as English and Malayalam may not have subject-verb agreement. For example, “avan vanthaan” and “avaL vanthaaL”, i.e the verb “vanthaan” or “vanthaaL” is decided by the subject “avan” or “avaL”. However, in English “came” is the common verb for both “he” or “she”. Also, due to variation in structure namely subject-verb-object (SVO) or subject-object-verb (SOV) of the languages, VP translation is a challenging task. Several researches have been reported [4, 3, 5, 14, 9, 10, 6] with various methodologies such as rule-based, phrase-based, statistical-based, machine learning and hybrid techniques for machine translation. Government of India released[1] a tool Sampark for performing machine translation among

[1] https://sampark.iiit.ac.in/sampark/web/index.php/content

Indian languages. Recently, Microsoft claims that developing deep neural network for Indian language translations brings more accuracy[2]. Further, developing methodology that performs VP translation between different language families such as Indo-Aryan, Indo-European and Dravidian is a difficult task. The shared task VPT-IL@FIRE2018 focuses on VP translations between different language families. The goal of VPT-IL@FIRE2018 task is to research and develop techniques to English-Tamil and Hindi-Tamil VP translations. VPT-IL@FIRE2018 is a shared Task on Verb Phrase Translation in English and Indian languages collocated with Forum for Information Retrieval Evaluation (FIRE-2018). This paper focuses on developing a methodology which does not require any linguistic knowledge that can translate VPs between any two languages of different families.

2 Proposed Methodology

A Sequence to Sequence (Seq2Seq) [11, 2] deep neural network is used in our approach for English-Tamil and Hindi-Tamil verb phrase translations. The steps used in our approach are given below.

- Extract English / Hindi VP sequences and Tamil VP input sequences from the given training data (English / Hindi and Tamil sentences) using the VP mapping information.
- Split the English / Hindi VP sequences and Tamil VP input sequences into training and development sets
- Determine vocabulary from both English / Hindi VP input sequences and Tamil VP input sequences.
- Build a deep neural network using Seq2Seq model with the layers namely embedding layer, encoding-decoding layer and projection layer with attention wrapper.
- Extract English / Hindi VP sequences from English / Hindi sentences of the test data
- Predict the Tamil VP output sequences for the English / Hindi VP sequences.
- Construct the Tamil VP output sequences into required output format.

The steps are detailed below.

2.1 Extraction of VP Sequences

The given text consists of parallel sentences in English and Tamil languages for Task 1 and parallel sentences in Hindi and Tamil for Task 2. The input sentences are tagged with sentence id and language information. Figure 1 shows the example parallel sentences for English and Tamil and Figure 2 shows the parallel sentences for Hindi and Tamil.

[2] https://news.microsoft.com/en-in/features/indian-language-translation-using-deep-neural-networks-announcement/

Fig. 1. English and Tamil Parallel Sentences.

```
<Sent Id=1 lang='en'>ENG:The General of the Chozha forces in Lanka at that time
was Kodumbalur Poodhi Vikrama Kesari .</Sent>
<Sent Id=2 lang='en'>ENG:They went and reported the matter to him .</Sent>
<Sent Id=3 lang='en'>ENG:The first day he had the rest he needed .</Sent>

<Sent Id=1 lang='ta'>&TAM:கொடும்பாளூர்ப் பெரிய வேளார் பூதி விக்கிரம கேசரி அச்சமயம் இலங்கைப் படையின்
சேநாதிபதியாக இருந்தார் .</Sent>
<Sent Id=2 lang='ta'>&TAM:அவரிடம் போய்ச் சொன்னார்கள் .</Sent>
<Sent Id=3 lang='ta'>&TAM:முதல்நாள் அவனுக்கு தேவைப்பட்ட ஓய்வு கிடைத்தது .</Sent>
```

Fig. 2. Hindi and Tamil Parallel Sentences.

```
<Sent Id=3 lang='hi'>इधर देखे हुए कुछ उपनिवेशों में लगभग 1600 जन तक रहते हैं .</Sent>
<Sent Id=4 lang='hi'>अरुणाचल प्रदेश भारत का एक उत्तर पूर्वी राज्य है .</Sent>
<Sent Id=5 lang='hi'>कुतुबमीनार भारत की सबसे ऊँची मीनार है .</Sent>

<Sent Id=3 lang='ta'>இங்கு காணப்படும் சில குடியேற்றங்களில் சுமார் 1600 பேர் வரை வசிக்கின்றனர் .</Sent>
<Sent Id=4 lang='ta'>அருணாச்சல பிரதேசம் இந்தியாவின் வடகிழக்கு பிரதேசம் ஆகும் .</Sent>
<Sent Id=5 lang='ta'>குதுப்மினார் இந்தியாவின் மிக உயர்ந்த தூண் ஆகும் .</Sent>
```

We have prepared the data in such a way that Seq2Seq deep learning algorithm may be applied. The English / Hindi VP input sequences and Tamil VP input sequences are constructed separately by extracting verb phrases from English / Hindi and Tamil sentences based on the VP mapping which consists of information namely sentence id, source language, target language, VP id, VP source information and VP target information. The VP source and target information consists of VP start position and length fields. The format of VP mapping is given in Figures 3 and 4.

Fig. 3. English-Tamil VP Mapping.

```
<vpInfo sentId='1' srcLang='en' tgtLang='ta' vpId='1' vp_src_info='59,3'
vp_tgt_info='92,9'>
<vpInfo sentId='2' srcLang='en' tgtLang='ta' vpId='2' vp_src_info='18,8'
vp_tgt_info='20,11'>
<vpInfo sentId='2' srcLang='en' tgtLang='ta' vpId='3' vp_src_info='9,4'
vp_tgt_info='13,6'>
<vpInfo sentId='3' srcLang='en' tgtLang='ta' vpId='4' vp_src_info='37,6'
vp_tgt_info='24,10'>
<vpInfo sentId='3' srcLang='en' tgtLang='ta' vpId='5' vp_src_info='21,3'
vp_tgt_info='41,9'>
```

The VP start position and length fields are used to extract the verb phrases present in sentences. For the above examples, the verb phrases are extracted as shown in Figures 5 and 6

Fig. 4. Hindi-Tamil VP Mapping.

```
<vpInfo sentId='3' srcLang='hi' tgtLang='ta' vpId='6' vp_src_info='47,8'
vp_tgt_info='59,13'>
<vpInfo sentId='3' srcLang='hi' tgtLang='ta' vpId='7' vp_src_info='4,8'
vp_tgt_info='6,10'>
<vpInfo sentId='4' srcLang='hi' tgtLang='ta' vpId='8' vp_src_info='45,2'
vp_tgt_info='50,5'>
<vpInfo sentId='5' srcLang='hi' tgtLang='ta' vpId='9' vp_src_info='35,2'
vp_tgt_info='42,5'>
```

Fig. 5. English and Tamil Verb Phrase.

was	இருந்தார்
reported	சொன்னார்கள்
went	போய்ச்
needed	தேவைப்பட்ட
had	கிடைத்தது

2.2 Model Building using Seq2Seq Model

We have adopted Neural Machine Translation (NMT) framework [8, 7] based on Seq2Seq model for VP translation task. Figure 7 shows the different layers used in deep neural network to build model for VP translation.

The verb phrases that are extracted using the previous step are given to the deep neural network. Sequence of layers namely embedding layer, encoder-decoder layer and projection layer are employed in the neural network to obtain Tamil VPs. We have determined the vocabulary for both English / Hindi VP input sequences (source input sequences) and Tamil VP input sequences (target input sequences). The source input sequences and the target input sequences are splitted into training sets and development sets. The English / Hindi VP input sequences with m words $x_1, x_2, ...x_m$ and Tamil VP input sequences with n words $y_1, y_2, ...y_n$ where m need not be equal to n are given to the embedding layer. The embedding layer learns weight vectors from the source input sequences and target input sequence based on their vocabulary. These vectors are given to multi-layer LSTM that performs encoding and decoding operations. We have used an attention mechanism [1, 7] to obtain an overall word alignment between the source and target sequences. The main idea of attention mechanism is to have direct connection between the source and target by paying attention to relevant source words (English / Hindi) as we translate into Tamil phrase. projection

Fig. 6. Hindi and Tamil Verb Phrases.

रहते हैं	வசிக்கின்றனர்
देखे हुए	காணப்படும்
है	ஆகும்
है	ஆகும்

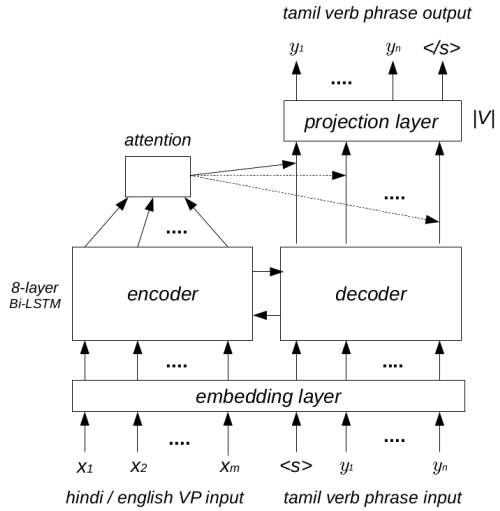

Fig. 7. System Architecture for Verb Phrase Translation.

layer that utilizes Softmax activation function is used to obtain the Tamil VP output sequences.

2.3 Prediction

The model that is built by using deep neural network is used to predict Tamil VP output sequences for the given English / Hindi verb phrases of test data. For this, we have extracted English / Hindi verb phrases from English / Hindi sentences of test data using the VP mapping information. The sample sentences given for English / Hindi languages and the corresponding VP mapping information for test data are shown in Figures 8 and 9.

Fig. 8. English and Hindi Sentences.

```
<Sent Id=1 lang='en'>The stone fell on Rishbaraja's head .</Sent>
<Sent Id=2 lang='en'>The bull shook the body one time .</Sent>
<Sent Id=3 lang='en'>It stared at the direction from which the stone had come .</
Sent>

<Sent Id=1 lang='hi'>वह कौन है , कैसे चेन्नई आया करके चलते है पहले हिस्सा .</Sent>
<Sent Id=2 lang='hi'>हिन्दू धर्म में केदारनाथ की आस्था और मान्यता का कोई पार नहीं है .</Sent>
<Sent Id=3 lang='hi'>वह कुंद की तराई के पूर्व भाग में प्रसिद्ध कुडुमियानमलै मंदिर स्थान पर स्थित है |</Sent>
```

The Tamil VP output sequences are obtained for the extracted English / Hindi VP input sequences using the deep neural model based on sequence mapping. We have constructed the Tamil VP output sequences for the test data into the required output format which is shown in Figure 10.

Fig. 9. VP Mapping for Test Data.

```
<vpInfo sentId='1' srcLang='en' tgtLang='ta' vpId='1' vp_src_info='10,4'>
<vpInfo sentId='2' srcLang='en' tgtLang='ta' vpId='2' vp_src_info='9,5'>
<vpInfo sentId='3' srcLang='en' tgtLang='ta' vpId='3' vp_src_info='3,6'>

<vpInfo sentId='1' srcLang='hi' tgtLang='ta' vpId='1' vp_src_info='24,3'>
<vpInfo sentId='1' srcLang='hi' tgtLang='ta' vpId='2' vp_src_info='33,4'>
<vpInfo sentId='2' srcLang='hi' tgtLang='ta' vpId='3' vp_src_info='56,7'>
```

Fig. 10. VP Translation Output.

```
<vpInfo sentId='1' srcLang='en' tgtLang='ta' vpId='1' vp_src_info='10,4'
translatedVP='விழுந்தான்'>
<vpInfo sentId='2' srcLang='en' tgtLang='ta' vpId='2' vp_src_info='9,5'
translatedVP='அல்ல'>
<vpInfo sentId='3' srcLang='en' tgtLang='ta' vpId='3' vp_src_info='3,6'
translatedVP='சொல்லிற்று'>

<vpInfo sentId='1' srcLang='hi' tgtLang='ta' vpId='1' vp_src_info='24,3'
translatedVP='பணித்தது'>
<vpInfo sentId='1' srcLang='hi' tgtLang='ta' vpId='2' vp_src_info='33,4'
translatedVP='எடுத்து'>
<vpInfo sentId='2' srcLang='hi' tgtLang='ta' vpId='3' vp_src_info='56,7'
translatedVP='இணைக்கும்'>
```

3 Implementation

We have used Python to extract VPs from English, Hindi and Tamil sentences. We have used TensorFlow for implementing the deep neural network. We have used the data set provided by VPTIL@FIRE2018 to evaluate our methodology. The data set used to evaluate the verb phrase translation task consists of a training set and test set for separately for English-Tamil and Hindi-Tamil. The training data contains parallel sentences in English and Tamil languages for Task 1 and parallel sentences in Hindi and Tamil languages for Task 2. VP mapping information is provided separately for both the tasks which consists of the attributes namely sentence id, source language, target language, VP id, VP source information and VP target information. The VP source and target information values have two parts namely VP start position and length of the VP. The details of the VPTIL@FIRE2018 data are given in Table 1.

Table 1. Data Set for VPTIL Task

Tasks	Training		Testing	
	No. of Sentences	No. of VPs	No. of Sentences	No. of VPs
English-Tamil	1443	2275	1096	1865
Hindi-Tamil	1992	2617	1000	1384

We have extracted the textual part of the input sentences by removing the <Sent> tags. For example, we have extracted the text “ENG:The General of the

Chozha forces in Lanka at that time was Kodumbalur Poodhi Vikrama Kesari ." from the input <Sent Id=1 lang='en'>ENG:The General of the Chozha forces in Lanka at that time was Kodumbalur Poodhi Vikrama Kesari .< /Sent> by removing <Sent Id=1 lang='en'> and < /Sent>. We have used the VP start position and length fields of VP source and target information from VP mapping to extract the verb phrases present in source and target languages. For example, the text that starts at position 59 for 3 character length is extracted from English with sentId 1 using the VP mapping information <vpInfo sentId='1' srcLang='en' tgtLang='ta' vpId='1' vp_src_info='59,3' vp_tgt_info='92,9'>. The obtained English VP input sequence with respect to vpId 1 is "was". For some sentences, the tokens for the verb phrase may not be continuous. For example, the VP mapping <vpInfo sentId='13' srcLang='en' tgtLang='ta' vpId='20' vp_src_info='41,7;56,14' vp_tgt_info='59,15'> conveys that the English VP sequence is present in two postions 41 and 56 with the length 7 and 14 respectively in the sentence <Sent Id=13 lang='en'>ENG:He opened his eyes and found the cat rubbing itself affectionately against him .< /Sent>. We have extracted the VPs in two positions as "rubbing" and "affectionately", and concatenated them as a single VP "rubbing affectionately" with respect to vpId 13 for English.

The extracted English / Hindi VP input sequences and Tamil VP input sequences are splitted into train set and development set to feed into the deep neural network. The details of the splits are given in Table 2.

Table 2. Number of Sequences for Model Building

Tasks	Training	Development
English-Tamil	1700	575
Hindi-Tamil	1817	800

We have used TensorFlow code based on tutorial code released by Neural Machine Translation [3] [7] that was developed based on Sequence-to-Sequence (Seq2Seq) models [11, 1, 8] to implement our deep learning approach for VP translations. We have implemented the Seq2Seq model using several parameters. The details are given below.

- Recurrent unit: LSTM
- Direction: Bi-directional
- No. of layers: 8
- Dropout: 0.2
- Batch size: 128
- Attention: Bahdanau
- Number of training steps: 50000

We have extracted the English / Hindi VP input sequences from the test data similar to training data. The Tamil VP output sequences are inferred with

[3] https://github.com/tensorflow/nmt

respect to the English / Hindi VP input sequences for the given test instances using our bi-LSTM model. Finally, we have converted the obtained Tamil VP output sequences into the required output format for the submission by adding the attribute as "translatedVP". The output format is shown in Figure 10.

4 Results

We have evaluated our models for English-Tamil and Hindi-Tamil VP translations using the data set provided by VPT-IL@FIRE2018 shared task. Table 3 shows the precision and recall values we have obtained for the test data using our models.

Table 3. Test Data Performance

Tasks	Precision(%)	Recall(%)
English-Tamil	10.06	16.53
Hindi-Tamil	16.84	18.21

It is observed from Table 3 that we have not obtained significant improvement in the performance. This is due to the size of the data set.

5 Conclusions

We have presented a deep learning approach based on Seq2Seq model for English-Tamil and Hindi-Tamil VP translations. We have used the data set provided by VPT-IL@FIRE2018 shared task. We have extracted English / Hindi VP sequences (source sequences) and Tamil VP input sequences (target sequences) from the given training data namely English / Hindi sentences and Tamil sentences respectively using verb phrase start position and length fields of source and target information present in the VP mapping file. These source and target input sequences are given to the deep neural network. The network consists of an embedding layer, encoding-decoding layer with 8-layer LSTM and a projection layer to translate the verb phrases from English / Hindi to Tamil. The embedding layer converts the source VP sequences and target VP input sequences into their vector representations based on the vocabulary of the source and target languages respectively. We have adopted Neural Machine Translation model for this task. The weight vectors learnt from embedding layer for training data are given to 8-layer LSTM where encoding and decoding are performed. We have used Bahdanau attention wrapper to obtain an overall word alignment between the source and target input sequences. Projection layer that uses Softmax activation function is used to obtain the Tamil verb phrase output sequences. This model is used to infer the Tamil VP output sequences for English / Hindi verb phrases of test data. Finally, the translated Tamil VP output sequences are converted to the required output format for submission. We have obtained precision

and recall values as 10.06% and 16.53% respectively for English - Tamil verb translations. For Hindi - Tamil verb translations, we have obtained precision and recall values as 16.84% and 18.21% respectively. The performance may be improved further with increased data set by incorporating more hidden layers, different attentions and increasing training steps.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
3. Devi, S.L., Pralayankar, P., Kavitha, V., Menaka, S.: Translation of hindi se to tamil in a mt system. In: Information Systems for Indian Languages, pp. 246–249. Springer (2011)
4. Devi, S.L., Pralayankar, P., Menaka, S., Bakiyavathi, T., Ram, R.V.S., Kavitha, V.: Verb transfer in a tamil to hindi machine translation system. In: Asian Language Processing (IALP), 2010 International Conference on. pp. 261–264. IEEE (2010)
5. Hasler, E., de Gispert, A., Stahlberg, F., Waite, A., Byrne, B.: Source sentence simplification for statistical machine translation. Computer Speech & Language **45**, 221–235 (2017)
6. Jadoon Khan, N., Anwar, W., Durrani, N.: Machine translation approaches and survey for indian languages. arXiv preprint arXiv:1701.04290 (2017)
7. Luong, M., Brevdo, E., Zhao, R.: Neural machine translation (seq2seq) tutorial. https://github.com/tensorflow/nmt (2017)
8. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
9. Macketanz, V., Avramidis, E., Burchardt, A., Helcl, J., Srivastava, A.: Machine translation: Phrase-based, rule-based and neural approaches with linguistic evaluation. Cybernetics and Information Technologies **17**(2), 28–43 (2017)
10. Sridhar, R., Sethuraman, P., Krishnakumar, K.: English to tamil machine translation system using universal networking language. Sādhanā **41**(6), 607–620 (2016)
11. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104–3112 (2014)
12. Thenmozhi, D., Aravindan, C.: Tamil-english cross lingual information retrieval system for agriculture society. In: Tamil Internet Conference (TIC2009), 9th International Conference on. pp. 173–178 (2009)
13. Thenmozhi, D., Aravindan, C.: Ontology-based tamil–english cross-lingual information retrieval system. Sādhanā **43**(10), 157:1–14 (2018)
14. Wang, X., Tu, Z., Xiong, D., Zhang, M.: Translating phrases in neural machine translation. arXiv preprint arXiv:1708.01980 (2017)