Overview of Verb Phrase Translation in Machine Translation: English to Tamil and Hindi to Tamil

Vijay Sundar Ram R and Sobha Lalitha Devi

AU-KBC Research Centre, MIT Campus of Anna University, Chennai, India
sobha@au-kbc.org

Abstract. We present an overview of verb phrase translation in machine translation from English to Tamil and Hindi to Tamil track, where English, Hindi and Tamil belong to three different language families, namely, Indo-European, Indo-Aryan and Dravidian family respectively. Verb phrases carry syntactic information such as tense, aspect, modal, and PNG (person, number and gender) other than the main verb. The characteristics of verb phrase vary between languages, which make the task challenging. In Tamil, non-finite verbs introduce clauses, whereas in Hindi clauses are introduced by relative-correlatives and the clause has a finite verb. We have five registrations and three out of the five registered teams submitted their runs. There were three submissions for English to Tamil Verb phrase translation and two submissions for Hindi to Tamil verb phrase translation. The runs are evaluated based on the correctness of tense, aspect, modal and PNG of the verb phrase.

Keywords: Verb phrase translation, English-Tamil, Hindi-Tamil.

1 Introduction

Machine translation (MT) is one of the most active research areas in Natural Language Processing across the globe. MT in Indian languages has picked-up in the last two decades. In developing a machine translation system, translation of the verb phrase from source language to target language is a challenging task. The verb phrases include finite verb, non-finite verb, auxiliary verb, main verb, verbal particles and negation verb constructions. Verb phrases also carry information namely, tense, aspect, modal, and PNG (person, number and gender) other than the main verb. The characteristics of the verbs vary between languages. In languages such as Tamil, Telugu, Kannada, Hindi, the subject and the finite verb of the sentence agree in PNG. In languages such as English and Malayalam, there is no agreement between the subject and finite verb. Languages vary in structure also such as SVO, SOV. These characteristics make the translation of Verb phrases from one language to another a difficult task

The objective of this shared task is to boost the research in Machine translation in Indian languages. We have narrowed down the scope of the track to translation of Verb Phrases from English to Tamil and Hindi to Tamil. These three languages are from different language families namely, Indo-European, Indo-Aryan and Dravidian. Sentence structures and characteristics of the verbs vary largely across these languages and make the task an interesting and challenging task. Sobha et al [1] has presented a work on transfer of verb phrases from Tamil to Hindi.

The researchers were welcomed to come-up with various methodologies such as rule-based, Machine Learning and Hybrid techniques. Participants were allowed to use any pre-processing tools, which are in open source or developed in-house.

The flow of the paper is as follows. In the following section, we have described about the data set provided for both the training and testing. We have given the details of the participants in the third section. In the fourth section, we have explained the methodologies followed by each team. The metrics for evaluation is presented in fifth section. The paper is concluded with a brief summary of the shared task.

2 Data set

We provided the training and testing dataset. The training data had a set of three files for each translation pair, which contains source language sentence, target language translated sentence and verb phase mapping index. Both the source and target language translated sentences has sentence indexing. Verb phrase (VP) mapping index file has the sentence index and the information of the position of the VP in the source language sentence and the position of the corresponding verb phrase in the target language translated sentence. The structure of the VP map index fie is as follows.

<vpInfo sentId='' srcLang='en/hi' tgtLang='ta' vpId='verbphrase-id' vp_src_info='' vp_tgt_info=''>

where:

sentId: is the sentence Id.

srcLang: Source language code. It can be 'en/hi'

tgtLang: Target language code. It is 'ta' as Tamil is the target language in both the pairs.

vpId: Each verb phrase is marked with an unique id.

vp_src_info: 'verb phrase start position and its length'

vp_tgt_info: 'verb phrase start position and its length'

We present below a sample source sentence, target translated sentence and its verb phrase map index from both English to Tamil and Hindi to Tamil in example 1 and 2 respectively.

Ex 1:
English to Tamil translated Sentence:

Source Sentence:
<Sent Id=24 lang='en'> Vandiyathevan did not get up .</Sent>

Translated Sentence:

<Sent Id=24 lang='ta'> வந்தியத்தேவன் எழுந்திருக்கவில்லை . </Sent>

VP Map Index:
<vpInfo sentId='24' srcLang='en' tgtLang='ta' vpId='36' vp_src_info='15,14' vp_tgt_info='16,18'>

The source English sentence has a verb phrase ‘did not get up’. And the translated Tamil sentence has the equivalent verb phrase ‘எழுந்திருக்கவில்லை’. vp_src_info in VP Map Index has the starting and length information of the VP in the source sentence, 15 and 14 respectively. Similarly vp_tgt_info has the starting and length information of the VP in the source sentence, 16 and 18 respectively. An unique Id is given to the verb phrase.

Ex 2:
Hindi to Tamil translated Sentence:

Source Sentence:

<Sent Id=8 lang='hi'>एथेंस महाद्वीप, सैलानयिों को रोमांचति कर देने वाला एक आकर्षक पर्यटक स्थल है .</Sent>

Translated Sentence

<Sent Id=8 lang='ta'>ஏதென்ஸ் பூகண்டம் சைலானியர்களை மெய்சிலிர்க்க வைக்கும் அழகான சுற்றுலாத் தலமாக உள்ளது .</Sent>

VP Map Index:
<vpInfo sentId='8' srcLang='hi' tgtLang='ta' vpId='12' vp_src_info='38,21' vp_tgt_info='30,22'>
<vpInfo sentId='8' srcLang='hi' tgtLang='ta' vpId='13' vp_src_info='82,2' vp_tgt_info='76,6'>

The source sentence, Hindi sentence, has two verbs ‘कर देने वाला’ and ‘है’, whose equivalent in Tamil are ‘வைக்கும்’ and ‘உள்ளது’ respectively. In VP map Index, the position of the verbs in both source and target sentences are given along with an unique id for each verb phrase.

In certain verb phrase such as verb phrase in interrogative sentence, the verbs occur separately with the words in-between. Consider the following example Ex 3. For these types of sentence, VP mapping index will have positional information of the verbs separated by “;”. An example is given below.

Ex 3.
Source Sentence:
<Sent Id=40 lang='en'>ENG:"How did that happen , Swami ?"</Sent>

Translated Sentence:
<Sent Id=40 lang='ta'>TAM:"அது எப்படி நடந்தது சுவாமிகளே!"</Sent>

VP Mapping Index:
<vpInfo sentId='40' srcLang='en' tgtLang='ta' vpId='58' vp_src_info='9,3;18,6' vp_tgt_info='17,7'>

Consider the above example, the source sentence, an integrative sentence, has the verb phrase 'did happen', occurring separately with 'that' in the middle. vp_src_info in VP Map Index has '9,3;18,6' its value. It has the starting position inform of both 'did' and 'happen' and their lengths.

The verb phrases include finite verb, non-finite verb, auxiliary verb main verb, verbal particles and negation verb constructions.

The statistics of the VPs indexed in the training and testing data is given in the table 1 below.

Table 1. Statistics of the training and testing data

S.No	Language Pair	Details	Training Data	Testing Data
1	Hindi to Tamil	Number of Sentences	1443	1098
		VPs indexed	2617	1384
2	English to Tamil	Number of Sentences	1992	1000
		VPs indexed	2267	1869

3 Data set Participants Details

We had five registrations. All participants registered for both English to Tamil VP translation and Hindi to Tamil VP translation tracks. Details of the participants are given in the following table 2.

Table 2. Registered Participants details

S.No	Institution	TeamId
1	Department of Computer Science, Banasthali Vidyapith, Banasthali - 304022, Rajasthan	Joshi-Banasthali
2	National Institute of Technology Mizoram, Chalatlang, Aizawl 796012, Mizoram, India	Pakray-NITM
3	IIIT-Delhi	Choudhary-IIITD
4	SSN College of Engineering, Old Mahabalipuram Road, Kalavakkam, Chennai	Thenmozhi-SSN

5	Centre for Applied Linguistics and Translation Studies, University of Hyderabad, Gachibowli Hyderabad	Parameswari-HCU

Out of five registered participants, three participants submitted their runs. The submission details are given in the table 3.

Table 3. Runs submitted Participants

S.No	Team	Runs submitted for language pairs
1	Choudhary-IIITD	English-Tamil and Hindi-Tamil
2	Thenmozhi-SSN	English-Tamil and Hindi-Tamil
3	Parameswari-HCU	English-Tamil

4 Methodologies

Choudhary-IIITD has used a neural machine translation technique using word-embedding along with Byte-Pair-Encoding (BPE) to develop an efficient translation system, called MIDAS translator. We used OpenNMT-py a neural machine translation toolkit for training our model. It is based on py-torch. They have used this translation system for both of the translation pairs i.e. English-Tamil and Tamil-Hindi.

Data pre-processing task included tokenization of sentences using their own tokenizer. The tokenization available in OpenNMT-py is not used.

After different trials they reported the best results using a Byte-pair-Encoded vocabulary , 2 Layer Bi-directional encoder-decoder, Adam optimization with a learning rate of 0.001, dropout (regularization) of 0.3, Bahdanau attention, and word-embedding with the dimension of 500.

Thenmozhi-SSN has adopted Neural Machine Translation model for this task. They have used a deep learning approach based on Seq2Seq model for English-Tamil and Hindi-Tamil VP translations. 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.

They have used TensorFlow for implementing the deep neural network.

They have implemented the Seq2Seq model using the bi-directional LSTM with 8 layers, dropout (regularization) of 0.2, Bahdanau attention, and the number of training steps is 50000.

Parashwari-HCU has used a rule-based approach for English to Tamil verb translation. They have performed the translation as three step process.

First they used a Stanford Dependency parser to identify the subject of the VP. The GNP information of the subject is noted. Second, using nltk lemmatizer the verb root is identified. Using a set of transfer rules the TAM of the target verb is generated and the equivalent Tamil for the verb root is replaced. In the third step, they have processed it with an in-house developed word-generator to generate Tamil verb phrase.

A brief summary of the methodology followed by the teams is presented in the table 4 below.

Table 4. Summary of Methodologies of each team

S.No	Team	Methodology	Tools
1	Choudhary-IIITD	A neural machine translation technique using word-embedding along with Byte-Pair-Encoding (BPE)	OpenNMT-py a neural machine translation toolkit
2	Thenmozhi-SSN	Neural Machine Translation model based on Seq2Seq model	TensorFlow
3	Parashwari-HCU	A rule based approach using a set of transfer rules	Stanford Depedency parser and nltk lemmatizer

5 Evaluation Methodology

We use a scoring methodology based on the correctness of the TAM (tense, aspect modal), Person, Number and Gender of the translated VP. The criteria for scoring the results are given in table 5.

Table 5. Scoring Criteria

Criteria	**Score**
Completely Correct	4
TAM and PNG Correct	3
Correct root and TAM partially correct	2
Correct root and wrong TAM	1
Completely Incorrect	0

English to Tamil Verb Phrase Translation

We had submissions from three teams for English to Tamil Verb phrase translation. We present teams against the Criteria Scores for English to Tamil in table 6. We have given details of number of verb phrases translation scored under each of the criteria scores for each of the teams. The precision and recall obtained by each team is presented in table 7.

Table 6. Team-wise Statistics of Verb phrase translations under each scoring criteria for English to Tamil VP Translation track

Teams	**Criteria Score**				
	4	**3**	**2**	**1**	**0**
Thenmozhi-SSN	107	9	104	89	1535
Choudhary-IIITD	347	18	229	116	388
Parashwari-HCU	282	7	152	100	287

Table 7. Team-wise Performance Measures for English to Tamil VP Translation track

Teams	Recall (%)	Precision (%)
Thenmozhi-SSN	16.53	10.06
Choudhary-IIITD	37.98	26.97
Parashwari-HCU	28.95	20.77

Hindi to Tamil Verb Phrase Translation

We had submissions from two teams for Hindi to Tamil Verb phrase translation. We present teams against the Criteria Scores for Hindi to Tamil in table 8. Similar to table 6, we have given details of number of verb phrases translation scored under each of the criteria scores for each of the teams in table 8. The precision and recall obtained by each team is presented in table 9.

Table 8. Team-wise Statistics of Verb phrase translations under each scoring criteria for Hindi to Tamil VP Translation track

Teams	Criteria Score				
	4	3	2	1	0
Thenmozhi-SSN	107	9	104	89	1535
Choudhary-IIITD	347	18	229	116	388
Parashwari-HCU	282	7	152	100	287

Table 9. Team-wise Performance Measures for Hindi to Tamil VP Translation track

Teams	Recall	Precision
Thenmozhi-SSN	18.21	16.84
Choudhary-IIITD	27.24	25.18

It is interesting to compare the number of verb phrase translations under each scoring criteria in table 6 and 8. We find more number of verb phrase translations under score 4 (Completely Correct) and score 2 (Correct root and TAM partially correct) and very less in criteria score 3 (TAM and PNG Correct). In both the tracks, all the submission shows that the generation of correct TAM and PNG is difficult. Identification of the root verb and translating it to the target language is found to be good based on the statistics of criteria score 2. In both Neural network based approach and transfer rule based approach the TAM and PNG generation to target language needs to be improved.

Conclusion

We have presented the overview of Verb phrase translation in English and Indian languages (VPT-IL) shared task. The shared task focused on translation of verb phrases from English to Tamil and Hindi to Tamil. As the three languages belong to three diffident language families and vary in their verb phrase formation, the task is a challenge. There are five registered teams, out of which three teams submitted their runs. Out of the three teams, two teams followed neural machine translation approach and the third linguistic rule based approach. Among the two neural machine translation, one used neural machine translation using word-embedding and the other used neural machine translation using Seq2Seq modeling. The third team with the rule-based approach used dependency parser to parse the source sentence and used a set of transfer rules to translate the verb phrases. The evaluation of the VP translation in both the tracks clearly presents the difficulties in generating correct TAM and PNG.

References

1. Sobha .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)