A Study of Neural Machine Translation Models

Abstract

Neural machine translation is a recently proposed framework for machine translation based purely on neural networks. The use of RNNs and LSTMs in machine translation has significantly improved the translation performance for longer sentences by being able to capture the context and long range correlations of the sentences in their hidden layers. The attention model based NMT system has become state-of-the-art, performing equal or better than other statistical MT approaches. In this project, we have studied the performance of different models for neural machine translation on the language pair, English and Hindi. We implemented the simple encoder-decoder model where the encoder encodes the input sentences into the annoted vectors which are used by the decoder to predict the next word in output target language based on the annoted vectors from the encoder, its own encoding of the previous state and the previous predicted word.

1 Introduction

Deep Neural Network has been successfully applied to machine translation.The work of (Cho et al., 2014; Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014) have shown that it is possible to build an end-to-end machine translation system using neural networks by introducing the *encoder-decoder* model. NMT systems have several advantages over the existing phrase-based statistical machine translation (SMT) systems (Koehn et al., 2007). The NMT systems do not assume any domain knowledge or linguistic features in source and target language sentences. Secondly, the entire encoder-decoder models are jointly trained to maximize the translation quality as opposed to the phrase-based SMT systems in which the individual components needs to be trained and tuned separately for optimal performance.

Although the NMT systems have several advantages, their performance is restricted in case of low-resource language pairs for which sufficiently large parallel corpora is not available and the language pairs whose syntaxes differ significantly. Morphological richness of language pairs poses another challenge for NMT systems that do not have any prior knowledge of the languages as it tends to increase the number of surface forms of the words due to inflectional attachments resulting in an increased vocabulary of the languages. Moreover, the inflectional forms have their semantic roles that have to be interpreted for proper translation. In order to enable the NMT systems to learn the roles of the inflectional forms automatically we need sufficiently large data. However, sufficiently large parallel data may not be available for low-resource morphologically rich language pairs. Most of the Indian languages are morphologically rich and there is lack of sufficiently large parallel corpus for Indian language paired with English. Given our familiarity with Hindi, we took up this task as a case-study and evaluated the performance of NMT models on language pair-English and Hindi.

2 NMT Models

2.1 Recurrent Neural Networks

Unlike the conventional translation models, where only a finite window of previous words would be considered for conditioning the language model, Recurrent Neural Networks (RNN) are capable of

conditioning the model on all previous words in the corpus. (Figure 1) introduces the RNN architecture where rectangular box is a hidden layer at a time-step, t. Each such layer holds a number of neurons, each of which performing a linear matrix operation on its inputs followed by a non-linear operation. At each time-step, the output of the previous step along with the next word vector in the document, $\mathbf{x}_t$, are inputs to the hidden layer to produce a prediction output $\hat{\mathbf{y}}$ and output features $\mathbf{h}_t$.

$$\mathbf{h}_t = \mathbf{W} f(\mathbf{h}_{t-1}) + \mathbf{W}^{(\mathbf{hx})} x_t \tag{1}$$

$$\hat{\mathbf{y}} = \mathbf{W}^{(\mathbf{S})} f(\mathbf{h}_t) \tag{2}$$

The amount of memory required to run a layer of RNN is proportional to the number of words in the

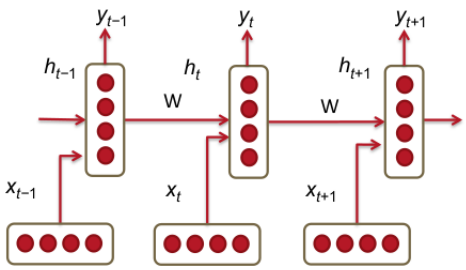

Figure 1: Recurrent Neural Network

corpus. For instance, a sentence with k words would have k word vectors to be stored in memory. Also, the RNN must maintain two pairs of W, b matrices. While the size of W could be very large, it does not scale with the size of the corpus (unlike the traditional language models). For a RNN with 1000 recurrent layers, the matrix would be 1000 x 1000 regardless of the corpus size.

2.2 Bi-directional RNN

It is possible to make predictions based on future words by having the RNN model read through the corpus backwards. (Irsoy et al.) shows a bi-directional deep neural network, that at each time-step,t, maintains two hidden layers, one for the left-to-right propagation and another for the right- to-left propagation. To maintain two hidden layers at any time, this network consumes twice as much memory space for its weight and bias parameters. The final classification result, $\hat{\mathbf{y}}_\mathbf{t}$, is generated through combining the score results produced by both RNN hidden layers.

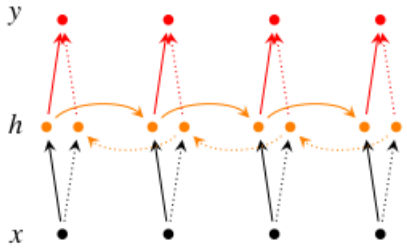

Figure 2: Bi-directional RNN

$$\overrightarrow{\mathbf{h}_t} = f(\overrightarrow{\mathbf{W}}\mathbf{x_t} + \overrightarrow{\mathbf{V}}\mathbf{h_{t-1}} + \overrightarrow{\mathbf{B}}) \tag{3}$$

$$\overleftarrow{\mathbf{h}_t} = f(\overleftarrow{\mathbf{W}}\mathbf{x_t} + \overleftarrow{\mathbf{V}}\mathbf{h_{t+1}} + \overleftarrow{\mathbf{B}}) \tag{4}$$

$$\hat{\mathbf{y}}_\mathbf{t} = g([\overrightarrow{\mathbf{h}_\mathbf{t}}; \overleftarrow{\mathbf{h}_\mathbf{t}}] + c) \tag{5}$$

2.3 Long-Short-Term-Memories(LSTMs)

In the mid-90s, a variation of recurrent net with so-called Long Short-Term Memory units, or LSTMs, was proposed by the German researchers Sepp Hochreiter and Juergen Schmidhuber as a solution to the vanishing gradient problem. The mathematical formulation of LSTM units are as follows :

$$i^{(\mathbf{t})} = \sigma(\mathbf{W}^{(\mathbf{i})}\mathbf{x}^{(\mathbf{t})} + \mathbf{U}^{(\mathbf{i})}\mathbf{h}^{(\mathbf{t-1})}) \tag{6}$$

$$f^{(\mathbf{t})} = \sigma(\mathbf{W}^{(\mathbf{f})}\mathbf{x}^{(\mathbf{t})} + \mathbf{U}^{(\mathbf{f})}\mathbf{h}^{(\mathbf{t-1})}) \tag{7}$$

$$o^{(\mathbf{t})} = \sigma(\mathbf{W}^{(\mathbf{o})}\mathbf{x}^{(\mathbf{i})} + \mathbf{U}^{(\mathbf{o})}\mathbf{h}^{(\mathbf{t-1})}) \tag{8}$$

$$\bar{c}^{(\mathbf{t})} = \sigma(\mathbf{W}^{(\mathbf{c})}\mathbf{x}^{(\mathbf{i})} + \mathbf{U}^{(\mathbf{c})}\mathbf{h}^{(\mathbf{t-1})}) \tag{9}$$

$$\bar{c}^{(\mathbf{t})} = \sigma(\mathbf{W}^{(\mathbf{c})}\mathbf{x}^{(\mathbf{i})} + \mathbf{U}^{(\mathbf{c})}\mathbf{h}^{(\mathbf{t-1})}) \tag{10}$$

Equations (6), (7), (8), (9), (10) are the input gate, forget gate, output gate, new memory cell and final memory respectively.

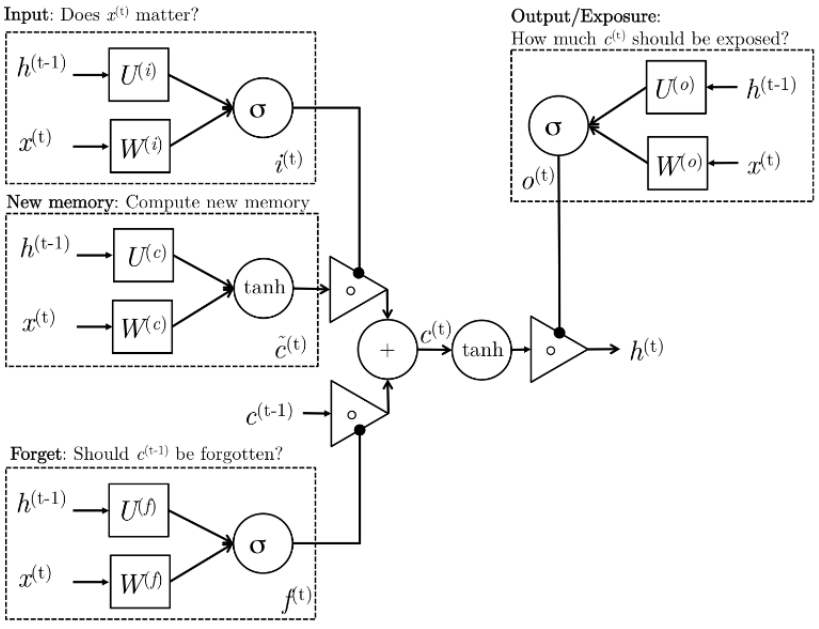

Figure 3: LSTM RNN

2.4 Attention Model

Bahdanau et al. (2014) and Luong et al. (2015) have proposed the attention-based translation model. The encoder of the model is a bi-directional RNN (Schuster and Paliwal, 1997). The annotation vectors $\mathbf{h_j^T}$ (where $\mathbf{h_j}$ encodes the j^{th} word with respect to the other words in the source sentence) are obtained by concatinating the two sequences of hidder layers $\overrightarrow{\mathbf{h_j^T}}$ and $\overleftarrow{\mathbf{h_j^T}}$ which are obtained by training the forward RNNs on the orginal sequence of input sentences and the backward RNNs on the reverse sequence of input sentences, such that $\mathbf{h_j^T} = [\overrightarrow{\mathbf{h_j^T}}; \overleftarrow{\mathbf{h_j^T}}]$. The decoder consists of a single layer GRU. At time step t, the alignment layers decides the relevance of the source words for the word to be predicted. The relevance (α_{tj}) of the j^{th} annotation vector at time t is determined by a feed-forward neural network that takes the previous state of the hidden layer of the decoder ($\mathbf{s}_{t-1}$), embedding of the last predicted word ($\mathbf{y}_{t-1}$) and the j^{th} annotation vector ($\mathbf{h}_j$) as input. The hidden state of the decoder at time t is computed as a function f_r of the previous hidden state $\mathbf{s}_t$, the context vector $\mathbf{c}_t$ and the previous predicted word $\mathbf{y}_{t-1}$,

where f_r is a GRU and $\mathbf{c}_t$ is the context vector for the t^{th} word is obtained as a sum of the annotation vectors weighted by the corresponding relevance scores.

$$\mathbf{s}_t = f_r(\mathbf{s}_{t-1}, y_{t-1}, c_t) \tag{11}$$

Finally, the conditional distribution over the words is obtained by using a deep output layer.

$$p(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{x}) \propto exp(\mathbf{y}_t^T(W_o f_o(\mathbf{s}_t, \mathbf{y}_{t-1}, \mathbf{c}_t) + b_o)) \tag{12}$$

where, $\mathbf{y}_t$ is the indicator vector corresponding to a word in the target vocabulary. W_o and b_o are the weights and bias of the deep layer and f_o is a single-layer feed-forward neural network with a two-way maxout layer (Goodfellow et al., 2013).

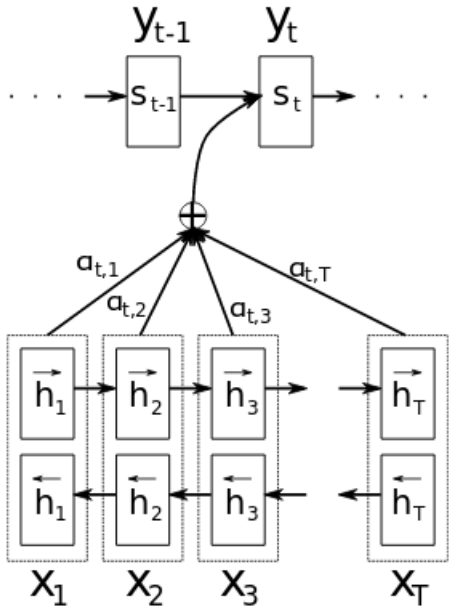

Figure 4: Attention model

Once the model learns the conditional distribution, then given a source sentence we can find a translation that approximately maximizes the conditional probability using, for instance, a beam search algorithm.

3 Proposed Work

In this project, we studied the performance of different NMT models on English-Hindi language pair. RNN's and LSTM's, can capture the context of the input seen so far in their hidden annoted vectors and hence perform better that the statistical translation models, and the results are more significant in case of longer sentences. The attention-based NMT models have shown near state-of-the-art performance for the language pairs, English-French and English-German. One of the main reason for the success was the availability of huge amount of training parallel corpus for these language pairs. But unfortunately not much of English-Hindi parallel corpus is available. One more thing to notice is that the sentence structure of English and Hindi is completely different. Ennglish follows the SVO sentence structure whereas Hindi follows a SOV sentence structure. We did an analysis of how the neural network models fare in case of language pairs that are structurally different and very less training corpus is available. We have implemented the simple encoder-decoder model, and for the underlying units, we used firstly a **Bi-RNN**, **LSTM** and then the **Attention Model** as proposed in (Bahdanau et al., 2014).

3.1 Resources used

Monolingual English and Hindi corpora were used to train word2vec (Mikolov et al., 2013) to obtain the word embeddings. The monolingual Hindi corpus was obtained from the ILTP-DC (www.tdil-dc.in/) which is about 10 GB. The English-Hindi parallel was obtained from WMT'14 which comprised of about 60,000 sentences obtained from different domains. From the 60000 English-Hindi parallel sentences about 59000 sentence pairs were randomly selected for training and remaining 1000 sentence pairs were used for testing. In order to reduce the size of the vocabulary we replaced all the numeric values by the 'NUM' token and the words in the vocabulary which had a frequency of less that two were replaced by the 'UNK' token.

The monolingual English wikipedia corpus of size 20.4 GB was used to obtain the English word vectors.

3.2 Implementation

The NMT models were implemented in Theano (Theano Development Team, 2016). The number of hidden layer units (n) were taken as 1000(in all the models), the word embedding dimensionality as 620 and the size of the maxout hidden layer in the deep output was 500. The number of hidden units in the alignment model(of the Attention based system) was 1000. We used gradient-clipping with a clipping threshold of 5. The model was trained using stochastic gradient descent with a learning rate of 0.0627 and batch size of 1. The model was run on a Nvidia Tesla K40C GPU machine.

4 Results

As the BLEU score suggests, the Attention model outforms the LSTM and the Bi-RNN models. Although in theory, the Bi-RNN and the LSTM models are able to capture the context of the input sentences in their hidden state vector, in practice they do not perform very well specially in case of longer sentences. We summarize the results and put six examples in tables in Appendix A, Appendix B and Appendix C for the Bi-RNN, LSTM and the Attention model respectively. One of the main reasons that can be attributed to the success of the attention model is the architecture used for the purpose. The (α_{ij}) parameter essentially emphasizes the importance of the j^{th} word in the source sentence in the translation of the i^{th} word in the target language sentence. Higher the value of (α_{ij}), more is the influence on the translation. And since we are using a Bi-RNN in the encoding of the source sentences, we ensure that the (α) parameters are trained of the input sequence in the forward as well as the backward direction. The single layer GRU used in the decoder takes into account its own previous context vector, the context vectors from the attention model and the previous predicted word, for predicting the next word.

Table 1: Comparison of 1) Bi-RNN model and 2) LSTM model and 3) Attention Model.

Translation model	**BLEU score**	**Iterations**
Bi-RNN translation model	11.10	33
LSTM translation model	15.07	15
Attention-based translation model	18.79	15

5 Analysis and Conclusion

Our Attention based model significantly outperformed the other two models and we went ahead and did further analysis of the results. We observed that our model did better on longer sentences but there were errors in the translation. On manually inspecting the translated sentences, we observed that the presence of named entities in the source language sentence was severely affecting our translation quality. Due to the limited size of the corpus, many named entities were absent from the vocabulary and hence the model was not able to find a suitable translation for them. It was also observed that certain phrases simply did not get translated into the target language. We observed their corresponding (α) values, and as expected, all the (α_{ij}) for those particular *j* which did not get translated were overall very less, indicative of the

fact that these source language words did not contribute much to the translation process. We take these problems as task for future scope work which we define in the next section as to how we can improve the quality of the translation.

6 Future Work

As mentioned in the previous section, we can take up as a further challenge how to improve the translation quality. The works of (Sudheshna et al.,2016) suggests two post processing heuristics that deal with named entities and untranslated words. (Jiajun Zhang et al.,2016) and (Birch et al.,2016) have shown that the parameters of the encoder and decoder can be fine-tuned separately using monolingual corpus of the source as well as the target languange. We are also looking into tree LSTMs in order to overcome the difficulties faced due to the structural differences in the source-target language pair.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron C. Courville, and Yoshua Bengio. 2013. Maxout networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 1319–1327.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. Seattle, October. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681, November.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Theano Development Team. 2016. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May.

Appendix A. Translation results of Bi-RNN model.

English	That drug taking could affect their health now and in the future.
Bi-RNN based TM	कि यिह बाहर और और और और और और और UNK ।
Ref.translation	कि ड्रग लेने से वर्तमान और आगे चलकर भवषि्य में उनके स्वास्थ्य पर इसका खराब असर पड़ सकता है ।
English	Like buddhism jainism flourished for some time and then began to decline.
Bi-RNN based TM	दक्षणि बौद्ध धर्म ही तथा के में में देश और में में में में ।
Ref.translation	बौद्ध धर्म की तरह जैन धर्म भी कुछ समय तक प्रगति करता रहा और बाद में उसका पतन प्रारंभ हो गया ।
English	I now went to a prison assignment as a police officer.
Bi-RNN based TM	मैं एक अपराध वशि्वविद्यालय ।
Ref.translation	मैं गयी थी एक पुलसि अधकिारी के तौर पे एक जेल के कार्य पे ।
English	Muslims believe that christians and jews have changed messages of their books.
Bi-RNN based TM	ऐसी से था था का ही और और और के के करता करता करता करता ।
Ref.translation	मुसलमान यह समझते हैं कि ईसाइयों और यहूदयिों ने अपनी पुस्तकों के संदेश में बदलाव कर दयि हैं ।
English	Ramayan Mahabharath and epics have been written in kavya sanskrit.
Bi-RNN based TM	रामायण के ही में ने तुलसी में का में गये ।
Ref.translation	रामायण महाभारत और पुराण काव्य संस्कृत में लखि गये हैं ।
English	The first batsman of the world making NUM runs in test cricket.
Bi-RNN based TM	क्रकिेट मुकाबले में दविसीय ज्यादा रन शतक से रन ।
Ref.translation	टेस्ट क्रकिेट में NUM रन बनने वाले वशि्व के पहले बल्लेबाज |

Appendix B. Translation results of LSTM model.

English	That drug taking could affect their health now and in the future.
LSTM based TM	जो कआिपका भवष्यि के हर UNK के तरीके से भी देख सकते हैं ।
Ref.translation	कड्रिग लेने से वर्तमान और आगे चलकर भवष्यि में उनके स्वास्थ्य पर इसका खराब असर पड़ सकता है ।
English	Like buddhism jainism flourished for some time and then began to decline.
LSTM based TM	जैसे बौद्ध धर्म कुछ जैन धर्म गरिावट तथा वकिास के बाद ।
Ref.translation	बौद्ध धर्म की तरह जैन धर्म भी कुछ समय तक प्रगतकिरता रहा और बाद में उसका पतन प्रारंभ हो गया ।
English	I now went to a prison assignment as a police officer.
LSTM based TM	अब मैं एक पुलसि अधकिारी के एक जेल गया एक जेल गया ।
Ref.translation	मैं गयी थी एक पुलसि अधकिारी के तौर पे एक जेल के कार्य पे ।
English	Muslims believe that christians and jews have changed messages of their books.
LSTM based TM	हम अपनी पुस्तक मुसलमान हैं ।
Ref.translation	मुसलमान यह समझते हैं कईसाइयों और यहूदयिों ने अपनी पुस्तकों के संदेश में बदलाव कर दयि हैं ।
English	Ramayan Mahabharath and epics have been written in kavya sanskrit.
LSTM based TM	वेद व्यास रामायण द्वारा प्राचीन वैदकि ग्रन्थ आदमिें हैं ।
Ref.translation	रामायण महाभारत और पुराण काव्य संस्कृत में लखि गये हैं ।
English	The first batsman of the world making NUM runs in test cricket.
LSTM based TM	वशि्व के प्रथम वशि्व की पहली बार NUM रन का NUM रन बनाने में ।
Ref.translation	टेस्ट क्रकिेट में NUM रन बनने वाले वशि्व के पहले बल्लेबाज |

Appendix C. Translation results of Attention based model.

English	That drug taking could affect their health now and in the future.
Attention based TM	कि्रग लेने से वर्तमान और भवषि्य में आगे और उनके स्वास्थ्य के भवषि्य में इसका असर होता है ।
Ref.translation	कि्रग लेने से वर्तमान और आगे चलकर भवषि्य में उनके स्वास्थ्य पर इसका खराब असर पड़ सकता है ।
English	Like buddhism jainism flourished for some time and then began to decline.
Attention based TM	कुछ समय के बाद कई वर्षों में जैन धर्म इसी तरह के बौद्ध धर्म की दशिा में वकिास और फरि गरिावट हैं ।
Ref.translation	बौद्ध धर्म की तरह जैन धर्म भी कुछ समय तक प्रगतकिरता रहा और बाद में उसका पतन प्रारंभ हो गया ।
English	I now went to a prison assignment as a police officer.
Attention based TM	अब मैं एक पुलसि अधकिारी की तरह एक जेल गया ।
Ref.translation	मैं गयी थी एक पुलसि अधकिारी के तौर पे एक जेल के कार्य पे ।
English	Muslims believe that christians and jews have changed messages of their books.
Attention based TM	मुसलमानों को यह मानते हैं कि्ईसाइयों और यहूदयिों के बारें में अपनी कतिाबों के बारें में बदल लयिा है ।
Ref.translation	मुसलमान यह समझते हैं कि्ईसाइयों और यहूदयिों ने अपनी पुस्तकों के संदेश में बदलाव कर दयि हैं ।
English	Ramayan Mahabharath and epics have been written in kavya sanskrit.
Attention based TM	रामायण महाभारत और बौद्ध का वर्णन भी संस्कृत में लखिा गया है ।
Ref.translation	रामायण महाभारत और पुराण काव्य संस्कृत में लखि गये हैं ।
English	The first batsman of the world making NUM runs in test cricket.
LSTM based TM	टेस्ट क्रकिेट NUM रन बनने वाले वशि्व के पहले बल्लेबाज ।
Ref.translation	टेस्ट क्रकिेट में NUM रन बनने वाले वशि्व के पहले बल्लेबाज |