Journal of Autonomous Intelligence (2019) Volume 2 Issue 4.
doi:10.32629/jai.v2i4.82

Original Article

A Study of Neural Machine Translation from Chinese to Urdu

Zeeshan Khan[1*], Muhammad Zakira[1], Wushour Slamu[1], Nady Slam[1]

School of Information Science and Engineering, Xinjiang University Urumqi, Xinjiang, China

ABSTRACT

Machine Translation (MT) is used for giving a translation from a source language to a target language. Machine translation simply translates text or speech from one language to another language, but this process is not sufficient to give the perfect translation of a text due to the requirement of identification of whole expressions and their direct counterparts. Neural Machine Translation (NMT) is one of the most standard machine translation methods, which has made great progress in the recent years especially in non-universal languages. However, local language translation software for other foreign languages is limited and needs improving. In this paper, the Chinese language is translated to the Urdu language with the help of Open Neural Machine Translation (OpenNMT) in Deep Learning. Firstly, a Chinese to Urdu language sentences datasets were established and supported with Seven million sentences. After that, these datasets were trained by using the Open Neural Machine Translation (OpenNMT) method. At the final stage, the translation was compared to the desired translation with the help of the Bleu Score Method.

Keywords: *Machine Translation; Neural Machine Translation; Non-Universal Languages; Chinese; Urdu; Deep Learning*

ARTICLE INFO
Received: Mar 10, 2020
Accepted: Apr 27, 2020
Available online: May 6, 2020

***CORRESPONDING AUTHOR**
Zeeshan Khan, School of Information Science and Engineering, Xinjiang University Urumqi, Xinjiang, China; zeeshan@uswat.edu.pk;

CITATION
Zeeshan Khan, Muhammad Zakira, Wushour Slamu, Nady Slam. A Study of Neural Machine Translation from Chinese to Urdu. Journal of Autonomous Intelligence 2019; 2(4): 29-36. doi: 10.32629/jai.v2i4.82

1. Introduction

In the era of globalization, communication, interaction, and cooperation among countries have become ever more common than before. At this period, failing to understand the foreign languages might impede the communication process. Machine translation (MT) is a helpful and efficient way to overcome barriers in communication between different languages. Machine translation has been making great progress in recent years, especially in the translation between universal languages such as English, German and French, English and Chinese, etc[1]. However, the number of local language translation software for non-universal languages is limited. Meanwhile, with the promotion of the “One Belt One Road “policy, many Chinese companies have initiated steps to reinforce mutual assistance between ASIAN countries. As one of the founding countries of ASIAN, Pakistan is also the largest economy in Southeast Asia. China and Pakistan have established cooperative relations in social infrastructure and trade. Therefore, the demand for translation in Chinese and Urdu is also increasing. With the help of machine translation, the difficulties that translators and interpreters have faced can be reduced and some difficulties in cross-language and cross-cultural cooperation can be eliminated, and this contributes to promoting political, economic and cultural exchanges between the two countries[2]. Machine translation (MT) is a subtype of computational linguistics that uses software to translate text from one natural language (NL) into another natural language (NL).

Machine translation executes simple translation of words from one language to another language, but this process is not a solution to give an accurate translation of a text due to the requirement of identification of whole phrases and their instant equivalents. To resolve this difficulty with corpus, statistical and neural techniques are the leading fields for better handling of differences in linguistic typology, isolation of anomalies, and translation systems. There are several types of machine translation (MT) systems; one of those is an Example-based Machine Translation or EBMT approach[3].

In recent years, Neural Machine Translation (NMT) has become one of the most popular machine translation methods.

1.1 Machine translation

MT device alongside in reality substituting the phrases as according to version also takes the application of complex linguistic understanding, morphology, meaning, and grammar into attention. The standard metric humans are using for evaluation of MT systems is the BLEU score.

Bilingual Assessment Understudy (BLEU) is the algorithm to define the superiority of text translated by a machine translation quality that is the assessment between machine-translated outputs to that of human-generated output; the closer machine translation is to human-generated translation, the better is the BLEU score. BLEU score is an n-gram overlap of gadget translation to that of reference translation[4]. As shown in (Eq. 1).

$$blue = min\left[1, \frac{output\ length}{reference\ length}\right]\left(\sum_{i=0}^{n} precision^{1/2}\right) \quad (1)$$

1.1.1 Direct machine translation (DMT)

Direct translation systems are basically bilingual and uni-directional. Direct translation approach needs only a little syntactic and semantic analysis. SL analysis is oriented specifically to the production of representations appropriate for one particular TL. DMT is a word-by-word translation approach with some simple grammatical adjustments[5].

1.1.2 Rule-Based machine translation (RBMT)

Rule based totally machine translation uses hand written linguistic regulations for both languages in its translation method. It calls for loads of human effort to outline the policies and adjustments of regulations generally costs very excessive[6]. It has 3 different sorts: (1) Direct; (2) Interlingua; (3) Transfer based.

In direct gadget translation, source language is immediately transformed to goal language without the use of intermediate steps. In Interlingua machine translation, there are intermediate steps which includes all necessary records for producing texts of target language. Interlingua steps generally layout in order to make it established for all pair of languages. In transfer based totally translation, there is bilingual representation of both languages in intermediate steps. This intermediate steps are language structured[7].

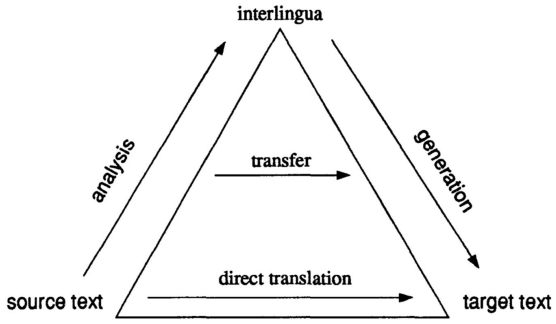

Figure 1. Types of rule based machine translation.

1.1.3 Interlingua-Based translation (IBT)

In IBT, the input language (IL) is converted into Interlingua form. The parser and analyzer of IL do not depend on a generator for the output language (OL). So, there is a condition for a complete determination of ambiguity in IL text.

1.1.4 Knowledge-Based machine translation (KBMT)

In KBMT, the source text and linguistic semantic information about the meaning of words and the combinations of words precedes the translation process into the target text. It is implemented over Interlingua architecture.

1.1.5 Statistical-Based machine translation (SBMT)

SBMT is a work on bilingual data. The SBMT is generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. The standard translation technique is established by the selection of the highest probability[8], as shown in (Eq. 2).

$$e = arg\,Max\,P\left(e/_f\right) = arg\,Max\,P\left(f/_e\right)P(e) \quad (2)$$

1.1.6 Example-Based machine translation (EBMT)

Example based system translation works on decomposing/fragmentation of supply sentence, translating these fragments into goal language and then re-composing the ones translated fragments into lengthy sentence[9].

1.1.7 Hybrid-Based machine translation (HBMT)

In hybrid machine translation, a combination of two or more machine translation techniques is used to overcome the limitations of each technique and to enhance the quality of translation[10].

1.1.8 Neural machine translation (NMT)

The neural networks and deep learning approach are used by this system. NMT models require only a fraction of the memory needed by old SMT. All parts of the NMT model are trained jointly, i.e., end to end, to make the best use of the translation performance. The NMT models trust on sequential encoder and decoder[11,12] without any unambiguous modeling of the syntactic structure of sentences. With this inspiration, the researchers made an effort to expand the translation model by modeling the hierarchical structure of language. Eriguchi first proposed a tree-based attentive NMT model[13], which was further extended by Yang[14] and Chen[15] via a bidirectional encoding mechanism. All the above tree-based models applied constituent tree structure and met the same difficulties. Other studies try to improve the NMT by modeling the syntax on the target side[16 – 19]. Regardless of enhancing the decoder, their success also proved the requirement and efficiency of modeling syntactic information for NMT systems.

2. Related Work

The machine translation method for language translation had started typically from 1990. There are several methods and approaches were developed in this field[3]. According to modern exploration, the performance report of baseline systems in translating Indian languages-based text (Bengali, Hindi, Malayalam, Punjabi, Tamil, Telugu, Gujarati, and Urdu) into English text has an average of 10% correctness for all language pairs[20].

In 2013 Kalchbrenner proposed recurrent continuous translation models for machine translation[21]. This model uses a Convolutional Neural Network (CNN) to encode a given part of input text into an unbroken vector and then uses a Recurrent Neural Network (RNN) as a decoder to convert the vector into output language.

In 2014, Long Short-term Memory (LSTM) was introduced into NMT[11]. To solve the problem of generating fixed-length vectors for encoders, they introduce attention mechanisms into NMT[12]. The attention mechanism allows the neural network to pay more consideration to the relevant parts of the input, and discard unrelated parts. Since then, the performance of the neural machine translation method has been significantly improved.

In this Sutskever a multilayer LSTM is used to encode input sentence into a fixed-size Direction and then decode it into output by another LSTM. The use of LSTM efficiently resolved the problem of gradient vanishing, which agrees on the model to capture data over extended space in a sentence. Muhammad Bilal uses the three classification models that are used for text classification using the Waikato Environment for Knowledge Analysis (WEKA). The blogs which are written in Roman Urdu and English can be considered as documents that are useful for the training dataset, labeled examples, and texting data. Due to testing, these three different models and the results were examined in each case.

The results display that Naive Bayesian outperformed Decision Tree and KNN in terms of more

accuracy, precision, recall, and measure[22]. Mehreen Alam addresses this difficult and convert Roman-Urdu to neural machine translation that guess sentences up to length 10 while achieving good BLEU score[23]. Neelam Mukhtar describes the Urdu language as a poor language that is mostly be ignored by the research community. After collecting data from many blogs of about 14 different genres, the data is being noted with the help of human annotators. Three well-known machine learning algorithms were used for the test and comparison: Support Vector Machine, Decision tree and k-Nearest Neighbor (k-NN).

It shows that k-NN performance is better than the Support Vector Machine and Decision tree in terms of accuracy, precision, recall, and f-measure[24]. Muhammad Usman also describes five well-known classification techniques on the Urdu language corpus. The corpus contains 21769 news documents of seven categories (Business, Entertainment, Culture, Health, Sports, and Weird). After preprocessing 93400 features that are taken out from the data to apply machine learning algorithms up to 94% precision[25]. In Yang and Dahl ' s work, firstly word is trained with a huge monolingual corpus, and then the word embedding is modified with bilingually in a context-depended DNN HMM framework. Word capturing lexical translation information and modeling context information are used to improve the word alignment performance. Unfortunately, the better word alignment result generated but cannot give significant performance an end-to-end SMT evaluation task[26].

To improve the SMT performance directly, Auli Urdu transliteration into sequence to sequence learning difficulty. The Urdu corpus was created and pass it to enhanced the neural network language model, in order to use both the source and target side information. In their work, not only the target word embedding is used as the input of the network, but also the current target word[27]. Liu suggests an improver neural network for SMT decoding[28]. Mikolov is firstly used to generate the source and target word inserting, which work on one hidden-layer neural network to get a translation confidence score[29]. The main factor that reveals the importance of this study is an absence of academic studies that conducted on the development of a Chinese and Urdu sentence-to-sentence translation model. This translation project is modeled as a neural machine translation and will make a significant contribution to the development of today's technological age.

3. OpenNMT

OpenNMT is an open-source tool that based on a neural machine translation system built upon the Torch/Py-Torch deep learning toolkit. The tool is designed to be user-friendly and easily accessible while also providing a high translation accuracy. This tool delivers a general-purpose interface, which needed only source and target data with speed as well as memory optimizations. OpenNMT has an active open public-friendly industrial as well as academic contribution. The diagram view of a neural machine translation is explained in **Figure 2**. The red source words are drawn to word vectors for a recurrent neural network (RNN). After finding the symbol <eos> then

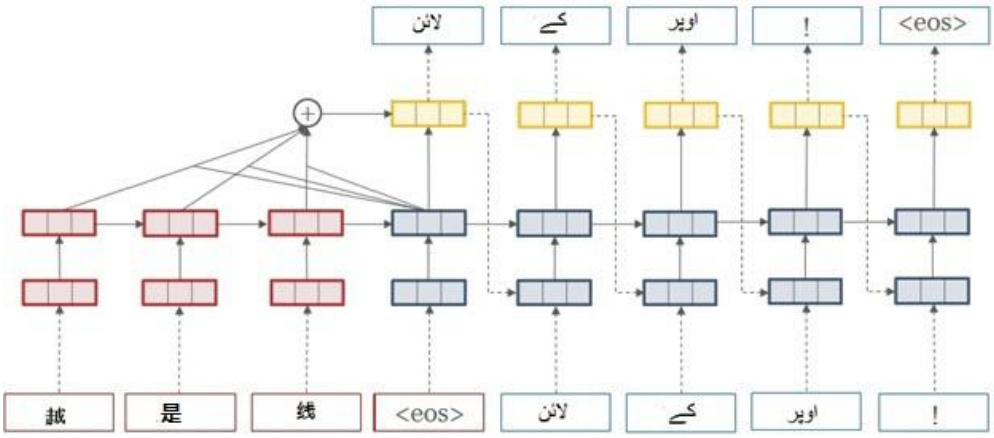

Figure 2. The view of NMT.

At the end of the sentence, the final step initializes a target blue RNN. At each step, the target is compared with source RNN and matched with the current hidden state, which shows prediction as declared in (Eq. 3). Due to this prediction, it provides for back into the target RNN.

$$P\left(wt/_{w1:_{t-1}},X\right) \quad (3)$$

Given model was trained with the help of OpenNMT Torch/PyTorch. There was no work on translation from Chinese to Urdu in the previous years. This study was conducted to reduce the barriers in the communication process between these two countries in the field of business and cultural promotion. Firstly, a Chinese-Urdu language parallel sentences datasets with more than million sentences were established. After that, these datasets were trained by using the Neural Machine Translation (NMT) method and traditional statistical machine translation.

3.1 Parallel corpuses

The amount of parallel corpus and its quality plays important role in quality of translation. For low resource languages like Urdu, it is extremely difficult to find sufficient parallel corpus for training, validation and testing of translation engine. The dataset of this project consists of two million Chinese-Urdu parallel corpus which derived from the combination of all the below datasets which are defined below.

(1) Monolingual Corpus: the Urdu corpus is around 95.4 million tokens distributed in and around different websites in which we have used 2.5 million for our approach. This corpus is a mix of sources such as News, Religion, Blogs, Literature, Science, Education, etc[30].

(2) IPC: The Indic Parallel Corpus is a collection of Wikipedia documents of six Indian sub-continent languages translated into English through crowdsourcing in the Amazon Mechanical Turk (MTurk) platform[31].

(3) UCBT: UCBT dataset is a parallel corpus of Urdu Chinese. UrduChineseCorp contain 5 million parallel sentences. It is not freely available for open research.

(4) CTUS: Chinese to Urdu Sentence dataset is a collection of different categories of sentences derived internet, news and books. In addition, dataset contains 50,000 sentences written manually and derived from UNHD (Urdu Nastaliq Handwritten Dataset) to meet a Chinese-Urdu parallel deficit. Which is shown in **Table 1**.

Table 1. Chinese to Urdu dataset

Number of Words	Number of Nouns	Number of Verbs	Number of Particles	Punctuation	Number of sentences
2553053	387957	436759	268950	178923	700000

3.2 Methodology

Several methods of OpenNMT tool are explained in the following subcategories.

3.2.1 Normalization

This method is used for smooth transformation on the source sequences to classify and keep some specific sequence into a single representation only for the translation process.

3.2.2 Tokenization

The tokenization is a process of splitting a sentence into pieces, each piece of a sentence is called tokens. OpenNMT uses a space-separated technique for tokenization.

3.2.3 Byte per encoding (BPE)

BPE or byte pair encoding is a script compression method which working on pattern substitution. In this work, the BPE model is developed on the basis of tokenized source data. For languages sharing an alphabet, understanding BPE on two or more involved languages increases the reliability of division and decreases the problem of insertion or deletion of characters when copying data[32].

3.2.4 Rearranged for tokenization

Due to the BPE model, the previous tokenized sentences were rearranged. There is an overview of the

case feature and joiner annotation. The case feature enhances extra features to the encoder which will be optimized for each label and then fed as extra source input alongside the word. On the target side, these features will be expected by the net. The decoder is then able to decode a sentence and annotate each word. On the other hand, by activating the joiner annotate symbol, the tokenization is reversible.

3.2.5 Preprocessing

In this method, the data is passing from preprocessing which can generate word vocabularies and balance data size, which is used for training.

3.2.6 Data training

In data training process, default OpenNMT encoder and decoder, LSTM layers, and RNN are taken. The research is based on open-source codebase[33]. This codebase is written in Python, using PyTorch, an open-source software library. For the NMT model, a single-layer LSTM with outstanding network connections is used as a good mechanism to train a translation model.

3.2.7 Data translation

In data translation the default OpenNMT is using binary translation method for creating an output translation file which comes from source and target language datasets.

4. Results and Discussions

OpenNMT is an open-source tool that based on a neural machine translation system built upon the Torch/Py-Torch deep learning toolkit. The tool is designed to be user-friendly and easily accessible while also providing a high translation accuracy.

In the OpenNMT model, the Chinese-Urdu language dataset with seven million parallel corpuses is trained. The validation part of a source and target have been taken as 25% of training corpus. For testing the model 15k randomly selected sentences from the corpus have been identified. Furthermore, these sentences underwent nine different tests. The results of each data-test have been collected and compared to both manual translation and translation model outputs. Additionally, the BLEU score for each data-test is calculated. The details of each data-test are shown below in **Figure 3** and **Table 2**.

Table 2. Data-Test result represented with BLEU score

Data-test	BLEU Score	System Information
Test1	0.0678	CPU@ 2.70 GHz, Intel (R) core (TM) I5-4700
Test2	0.0847	
Test3	0.0855	
Test4	0.0889	
Test5	0.0924	
Test6	0.0929	
Test7	0.0987	
Test8	0.1156	
Test9	0.1887	

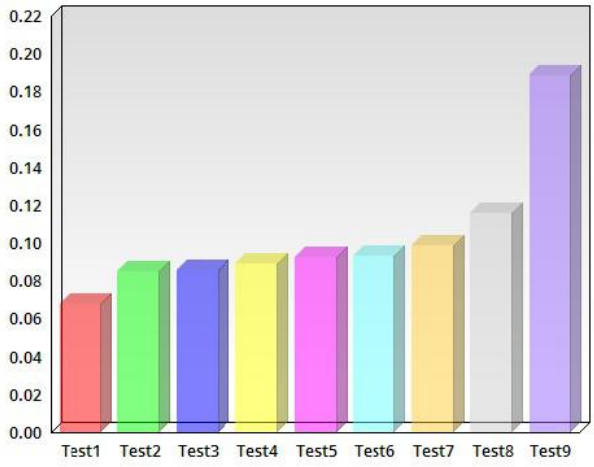

Figure 3. Histogram representation of BLEU score of different data-test.

5. Conclusions

This study consists of trained data from several sources that have been made up of different variety of sentences. The training part of method is conducted in the shape of data-test, and it has proved practical as the BLEU score has been increased with the number of data-test; the accuracy of the system is obtained after ninth data-test, which is suitably matched to other machine translation systems. The BLEU score of Chinese to the Urdu translation system is going to be improved by applying some more techniques, which are used to generate the best model of translation.

References

1. Jonathan Slocum. A survey of machine translation: Its history, current status, and future prospects. 1985; 11(1): 1-17.
2. Bai L, Liu W. A Practice on Neural Machine Translation from Indonesian to Chinese. Recent Trends in Intelligent Computing, Communication and Devices 2020; 33-38.
3. Godase A, Govilkar S. Machine translation development for Indian languages and its approaches. Behavioral & Brain Sciences 2015; 4(2): 55-74.
4. Papineni K, Roukos S, Ward T, *et al.* BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics 2002; 311-318.
5. Okpor M. Machine translation approaches: Issues and challenges. International Journal of Computer Science Issues 2014; 11(5): 159.
6. Mall S. and Jaiswal U. Survey: Machine translation for Indian language. 2018; 13(1): 202-209.
7. Hutchins WJ, Somers HL. An introduction to machine translation. Academic Press London 1992; Vol. 362.
8. Marcu D, Wong D. A phrase-based, joint probability model for statistical machine translation. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing 2002.
9. Zafar M, Masood A. Interactive English to Urdu machine translation using example-based approach. International Journal on Computer Science & Engineering 2009; 1(3): 275-282.
10. Pathak AK, Acharya P, Balabantaray RC. A case study of Hindi – English example-based machine translation. Innovations in Soft Computing and Information Technology 2019; 7-16.
11. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems 2014.
12. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. Computer Science 2014.
13. Eriguchi A, Hashimoto K, Tsuruoka Y. Tree-to-sequence attentional neural machine translation. 2016.
14. Yang B, Wong DF, Xiao T, *et al.* Towards bidirectional hierarchical representations for attention-based neural machine translation. 2017.
15. Chen H, Huang S, Chiang D, *et al.* Improved neural machine translation with a syntax-aware encoder and decoder. 2017.
16. Wu S, Zhang D, Yang N, *et al.* Sequence-to-dependency neural machine translation. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics 2017; Vol. 1.
17. Eriguchi A, Tsuruoka Y, Cho K. Learning to parse and translate improves neural machine translation. 2017.
18. Aharoni R, Goldberg Y. Towards string-to-tree neural machine translation. 2017.
19. Du W, Black AW. Top-down structurally-constrained neural response generation with lexicalized probabilistic context-free grammar. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2019; Vol. 1.
20. Khan NJ, Anwar W, Durrani N. Machine translation approaches and survey for Indian languages. 2017.
21. Kalchbrenner N, Blunsom P. Recurrent continuous translation models. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing 2013.
22. Bilal M, Israr H, Shahid M, *et al.* Sentiment classification of Roman-Urdu opinions using Naive Bayesian, Decision Tree and KNN classification techniques. Journal of King Saud University Computer & Information Sciences 2016; 28(3): 330-344.
23. Alam M, Hussain S. Sequence to sequence networks for Roman-Urdu to Urdu transliteration. International Multi-topic Conference (INMIC) 2017. IEEE.
24. Mukhtar N, Khan MA. Urdu sentiment analysis using supervised machine learning approach. International Journal of Pattern Recognition and Artificial Intelligence 2018; 32(2): 1851001.

25. Usman M. Urdu text classification using majority voting. 2016; 7(8): 265-273.
26. Yang N, Liu S, Li M, *et al.* Word alignment modeling with context dependent deep neural network. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics 2013; Vol. 1.
27. Auli M, Galley M, Quirk C, *et al.* Joint language and translation modeling with recurrent neural networks. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing 2013; 1044-1054.
28. Liu L, Taro W, Eiichiro S, *et al.* Additive neural networks for statistical machine translation. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics 2013; Vol. 1.
29. Mikolov T, Karafiat M, Burget L, *et al.* Recurrent neural network based language model. Eleventh Annual Conference of the International Speech Communication Association 2010.
30. Post M, Callison-Burch C, Osborne M. Constructing parallel corpora for six Indian languages via crowdsourcing. Proceedings of the Seventh Workshop on Statistical Machine Translation 2012; 401-409.
31. Baker P, Hardie A, McEnery T, *et al.* EMILLE, a 67-million word corpus of Indic languages: data collection, mark-up and harmonisation. Proceedings of the Third International Conference on Language Resources and Evaluation (LREC’02) 2002.
32. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. 2015.
33. Luong T, Brevdo E, Zhao R. Neural machine translation (seq2seq) tutorial. 2017.