Journal of Autonomous Intelligence (2021) Volume 4 Issue 1.
doi:10.32629/jai.v4i1.359

Original Article

A Seq to Seq Machine Translation from Urdu to Chinese

Zeeshan[1]*, Jawad[1], Muhammad Zakira[1], Muhammad Niaz[1]

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

ABSTRACT

Machine translation (MT) is a subtype of computational linguistics that uses to implement the translation between different natural languages (NL). Simply word to word exchanging on machine translation is not enough to give desire result. Neural machine translation is one of the standard methods of machine learning which make a huge improvement in recent time especially in local and some national languages. However these languages translation are not enough and need to focus on it. In this research we translate Urdu to Chinese language with the help of neural machine translation (NMT) in deep learning methods. First we build a monolingual corpus of Urdu and Chinese languages, after that we train our model using neural machine translation (NMT) and then compare the data-test result to accurate translation with the help of BLEU score method.

Keywords: *Machine Translation; Deep Learning; Neural Machine Translation; Urdu Language; Chinese Language*

ARTICLE INFO
Received: Jun 23, 2021
Accepted: Sep 24, 2021
Available online: Sep 24, 2021

***CORRESPONDING AUTHOR**
Zeeshan, School of Information Science and Engineering, Xinjiang University Urumqi, Xinjiang, China;
zeshanali531@gmail.com;

CITATION
Zeeshan, Jawad, Muhammad Zakira, Muhammad Niaz. A seq to seq machine translation from Urdu to Chinese. Journal of Autonomous Intelligence 2021; 4(1): 1-5. doi: 10.32629/jai.v4i1.359

1. Introduction

In the modern globalization era the communication between different countries are more frequent and important than before. So at the same time it's too much difficult to contact in different languages. Around the world there are almost 6500 languages, Language is one of the most powerful tools of any living being to convey their thoughts to the other but it is only possible if the communicating subjects have the same language. A language can be expressed as a series of spoken sounds and words or gestures. It's not possible for someone to learn or speak whole languages. So for this problem, researchers of computer science are interested in developing systems to improve the interaction between humans and computers to make communication possible between different countries[1]. In this field many exports did a lot of efforts on different tool techniques for convey our massage to each other between different languages smokers. In this paper we used Open Neural Machine translation Open (NMT), it is supportive and resourceful way to overcome the barricades in contact in different languages. Neural Machine translation has made great progress nowadays in translation between universal languages such as English and French, English and Chinese etc. However, the number of domestic translation software for non-universal languages is limited[2]. Recently, with the renaissance of deep learning, end-to-end Neural Machine Translation (NMT)[3], has gained incredible performance[4]. Early NMT solutions are typically optimized to maximize the chances of estimation (MLE) of each sentence in the ground accurate translations during the training processing time. However, such an objective cannot guarantee the sufficiency of the generated translations in the NMT model, due to the lack of quantitative measurement for the information transformational completeness from the source side to the target side.

2. Related Works

The Machine interpretation Technique for language interpretation and the other way around had begun extensive back regularly from 1990 onwards. The few techniques and their approaches[5]. According to current investigation, the presentation report of standard frameworks is translating Indian dialects based content (Bengali, Hindi, Malayalam, Punjabi, Tamil, Telugu, Gujarati, and Urdu) into English content with a normal of 10% Rightness for all language pairs[6]. In 2013 Kalchbrenner proposed intermittent persistent interpretation models for machine translation[2]. This model uses a convolution neural system (CNN) to encode a given piece of info content into a solid vector and afterward utilizes a recurrent neural organize (RNN) as a decoder to change over the vector into yield language. In 2014, long momentary memory (LSTM) was brought into NMT[7]. To take care of the issue of creating fixed-length vectors for encoders, they bring consideration component into NMT[3]. The consideration component permits then neural system to pay more thought to the significant pieces of the info, and dispose of in consequential parts. From that point forward, the presentation of the neural machine translation strategy has been essentially improved. In this Sutskever a multi-layer LSTM is utilized to encode input sentence into a fixed-size Bearing and afterward translate it into yield by another LSTM. The utilization of LSTM Proficiently settled the issue of inclination disappearing, which concurs the model to catch information over broadened space in a sentence. Muhammad Bilal utilizes the three order models are utilized for content grouping utilizing Waikato Condition for Information Examination (WEKA). Opinions written in Roman Urdu and English for blog. These suppositions are reports which are utilized for preparing data-set, named models and messaging information. Because of testing these three unique models and the outcomes for each situation are examined. The outcomes show that Gullible Bayesian outflanked Choice Tree and KNN as far as more exactness, accuracy, review and Measure[8]. Mehreen Alam address this troublesome and convert Roman-Urdu to Urdu transliteration into arrangement to grouping learning trouble. The Urdu corpus was make and pass it to neural machine interpretation that theory sentences up to length 10 while accomplishing great BLEU score[9]. Neelam Mukhtar depict Urdu language is poor dialects, for example, Urdu are generally disregarded by the examination network. In the wake of gathering information from numerous web journals of around 14 unique classes, the information is being noted with the assistance of human annotators. Three notable AI calculation Bolster Vector Machine, Choice tree and k-Closest Neighbor (k-NN) which is utilized for test, comparison. Its show that KNN execution is superior to Help Vector Machine and Choice tree as far as exactness, accuracy, review and f-measure[10]. Muhammad Usman additionally portray five notable order strategies on Urdu language corpus. The corpus contains 21769 news reports of seven classes (Business, Diversion, Culture, Well-being, Sports, and Odd). In the wake of preprocessing 93400 highlights are take out from the information to apply AI calculations up to 94% precision[11]. Yang and Dahl their work, first word prepared with a gigantic mono-lingual corpus, at that point the word installing is changed with bilingually in a setting depended DNN Well system. Word catching lexical interpretation data and demonstrating setting data for improve the word arrangement execution. Sadly, the better word arrangement result produced yet can't give critical execution an end-to-end SMT assessment task[12]. To improve the SMT execution straight forwardly Auli upgraded the neural system language model, so as to utilize both the source and target side data. In their work, not just the objective word implanting is utilized as the contribution of the system, yet in addition the present objective word[13]. Liu propose an improver neural system for SMT decoding[14]. Mikolov is right off the bat used to produce the source and target word embeddings, which take a shot at one covered up layer neural system to get an interpretation certainty score[15]. Due to the past inquires about, we convey a training on machine translation from Chinese to Urdu language.

3. Open NMT

Open NMT is an open-source device which dependent on neural machine interpretation framework based upon the Torch/Py-Torch deep learning toolbox. The apparatuses are intended to be easy to understand and effectively open while additionally giving high translation accuracy. This device conveys broadly useful interface, which required just source and target information with speed just as memory enhancements. Open NMT has dynamic open neighborly mechanical just as scholastic commitments. We train our model with the assistance of Open NMT torch/pytorch. There is no work done on Urdu to Chinese language interpretation in past years. We accomplished this work for evacuating challenges in correspondence between these two nations in business just as culture advance line. First we build up an Urdu to Chinese language equal sentences datasets which have in excess of parallel Sentences. From that point forward, we train Our Datasets utilizing open neural machine interpretation (Open NMT) strategy, with customary, measurable machine interpretation.

4. Parallel Corpus

Our dataset consists of 50k Urdu- Chinese parallel corpus which is come from the combination of all the below datasets which are define below.

Monolingual Corpus: which is collected from different Website in which Urdu corpus is around 95.4 5 tokens. These corpus is a combination of quantities such as News, Religion, Blogs, Literature, Science, Education etc.[16].

IPC: The Indic Parallel Corpus is a collection of Wikipedia documents of six Indian sub-continent languages translated into English through crowd sourcing in the Amazon Mechanical Turk (M-Turk) platform[17].

UTCS: Urdu to Chinese Sentence dataset is the collection of different group of sentences from different part of internet[1], news, some from manually hand write because a less parallel Urdu to Chinese sentences present over internet so due to not accessibility of parallel Urdu to Chinese data. We make our own datasets which size is in 50k with the help of taking some part from above datasets and some manually to make our parallel UTCS Dataset for training which is showing in **Table 1**.

Table 1. Urdu to Chinese dataset

Number of Words	Number of Nouns	Number of Verbs	Number of Particles	Punctuation	Number of Sentences
185305	48795	33675	26795	18892	500000

5. Experiment

The several procedures of Open NMT tool are explained in the following subsections.

6. Preprocessing

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

7. Data Training

We are selecting default Open NMT encoder and decoder, LSTM layers, and RNN. We start our research by using open-source code-base[18]. This code-base is written in Python, using pytorch, an open-source software library. We used two-layer LSTM with outstanding network connections as well as good mechanism to train a translation for our NMT model.

8. Data Translation

In data translation the Open NMT model using binary translation method for creating an output translation file which come from source and target language datasets.

9. Results and Discussions

In Open NMT model we are trained Urdu to Chinese language dataset which is 50k parallel corpus. The validation part of a source and target which have been taken as 20% of training corpus, then for testing the model we have taken 10k sentences randomly from the corpus then we make 7 different test in our model. We selected different result for each data-test and given name UCT and also compare the translation with manually as well as in translation model. We also calculated the BLEU score for every UCT. The details of the each data-test below in **Figure 1** and **Table 2**.

Table 2. UCT result represented with BLEU score

Data-test	BLEU Score	System Information
UCT 1	0.0678	CPU@ 2.70 GHz, Intel (R) core (TM) I7-5700
UCT 2	0.0847	
UCT 3	0.1089	
UCT 4	0.1929	
UCT 5	0.2187	
UCT 6	0.3856	
UCT 7	0.4287	

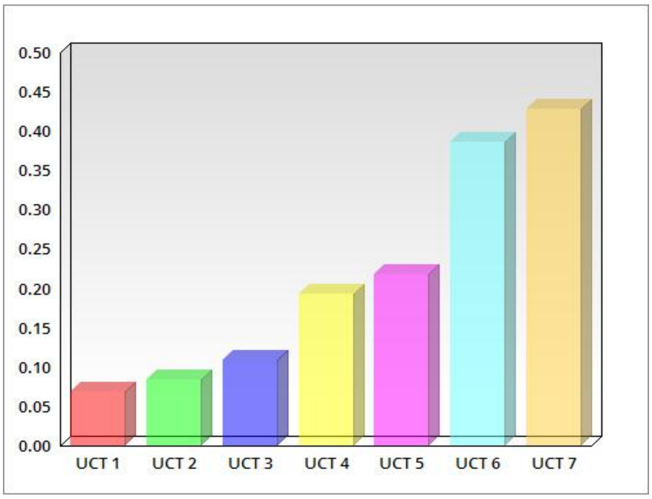

Figure 1. Bar representation of BLEU score of different UCT.

10. Conclusions

We have taken training data from several sources, which is to be made up of different variety of sentences. The training part of our method is done in the shape of UCT, and it has been practical that the BLEU score increases with the number of UCT; accuracy of the system obtained after seven number of UCT which is proper matched to other machine translation systems. We are still trying to improve the BLEU score of Urdu to Chinese translation system by applying some more techniques which are used for generating the best model of translation.

References

[1] http://www.statmt.org/wmt16/translation-task.html

1. Damerau FJ. A technique for computer detection and correction of spelling errors. Communications of the ACM 1964; 7(3): 171-176.
2. Kalchbrenner N, Blunsom P. Recurrent continuous translation models. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing 2013.
3. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. Preprint arXiv: 1409.0473, 2014.
4. Vaswani A, *et al.* Tensor 2 tensor for neural machine translation. Preprint arXiv: 1803.07416, 2018.
5. Godase A, Govilkar S. Machine translation development for Indian languages and its approaches. International Journal on Natural Language Computing(IJNLC) 2015; 4(2): 55-74.
6. Khan NJ, Anwar W, Durrani N. Machine translation approaches and survey for indian languages. Preprint arXiv: 1701.04290, 2017.
7. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems 2014.
8. Bilal M, *et al.* Sentiment classification of Roman-Urdu opinions using Naïve Bayesian. Decision Tree and KNN Classification Techniques 2016; 28(3): 330-344.
9. Alam M, Hussain Sibt ul. Sequence to sequence networks for Roman-Urdu to Urdu transliteration. In 2017 International Multi-topic Conference (INMIC) 2017. IEEE.
10. Mukhtar, Neelam, Khan, *et al.* Urdu sentiment analysis using supervised machine learning approach. International Journal of Pattern Recognition & Artificial Intelligence 2018; 32(2): 1851001.
11. Usman M, *et al.* Urdu text classification using majority voting. 2016; 7(8): 265-273.
12. Yang N, *et al.* Word alignment modeling with context dependent deep neural network. In Proceedings of the 51_{st} Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2013.
13. Auli M, *et al.* Joint language and translation modeling with recurrent neural networks. 2013.
14. Liu L, *et al.* Additive neural networks for statistical machine translation. In Proceedings of the 51_{st} Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2013.

15. Mikolov T, *et al.* Recurrent neural network based language model. In Eleventh Annual Conference of the International Speech Communication Association 2010.
16. Post M, Callison-Burch C, Osborne M. Constructing parallel corpora for six Indian languages via crowd-sourcing. In Proceedings of the Seventh Workshop on Statistical Machine Translation 2012. Association for Computational Linguistics.
17. Baker P, *et al.* EMILLE, A 67-Million Word Corpus of Indic Languages: Data Collection, Mark-up and Harmonisation. In LREC 2002.
18. Thang Luong, Eugene Brevdo, Rui Zhao. Neural machine translation (seq2seq) tutorial. Google Research Blogpost 2017.