Southern University of Science and Technology

CS322

Innovative Experiment II

Deep NMT Off-line Deployment in Mobile Devices

Author:	*Student Number:*
Yilin Zheng	11510506
Shupei Chen	11510319
Chenxuan Wang	11510488
Mentor:	*E-mail:*
Ke Tang	tangk3@sustc.edu.cn

April 12, 2018

Contents

1 Abstract

Machine translation is a hot field in computer science and statistics. After neural networks become popular studied again, this field is facilitated when neural network models are applied and the performance surprise all the researchers. However, fewer studies are conducted to find a suitable model which can run off-line on mobile devices compared to the approaches of a wide range of NMT services. This project will try to find and construct a deep neural network model running off-line and integrates the model into an Android application.

2 Introduction

This project is a research project on deep neural machine translation(NMT) systems off-line deployment in mobile devices, which is conducted by Yilin Zheng, Shupei Chen, and Chenxuan Wang, advised by Ke Tang in the department of Computer Science and Engineering, Southern University of Science and Technology. This project is not a one-semester long course project but a more academic research thesis.

2.1 Background

Nowadays, researchers have achieve a prominent effects on neural machine translation and speech recognition via deep-learning based models. However, such services are usually running large deep neural networks on large-scale computer clusters and are provided to users through the Internet. Therefore, when the Internet suffers bad conditions, the translation services won't work normally and users can only wait for Internet shifting to be stable. To avoid such awkward situation, researchers are considering the off-line deployment of NMT systems. However, compared to large computer clusters, mobile devices only provide lower computation and slower processing speed, which leads original complex deep neural networks infeasible to function. This project aims to design a feasible deep neural networks based on the state-of-art for NMT in mobile devices.

2.2 Project Basis

Based on the situation analysed in the last section, to solve the limits of Internet condition and improve the NMT service, this project is crucial and potential. Besides, neural network compression has been widely studied around the world which can provide feasible techniques to adapt certain deep neural network models to mobile devices. Recently, the hardware of mobile devices has made continuous upgrade so that the mobile devices can provide more powerful computation and faster processing speed. Some companies have developed mobile deep learning chips. For software, some enterprises provide deep learning toolkits such as Google's TensorFlow Lite, Apple's CoreML, Android's Neural Networks API, and Baidu's mobile-deep-learning. This project has a great feasibility inspired by these conditions.

3 Survey

This part is a survey of NMT researches. This part will roughly introduce the problem of machine translation, the task of NMT, some typical DNN models and techniques of NMT, related datasets and a universally used evaluation.

3.1 Problem

Machine translation is a sub-field of computational linguistics involving the software used to translate a language to another. Usually, the results of machine translation are worse than human translation since the difficulty of modelling the grammar and semantics.

The process of machine translation can be divided into two steps: one is encoding the input sentence and the other is decoding the processed source into the target language. The input is usually text or sentences $X = \{x_1, x_2, ... \cdots, x_n\}$ and the output can be represented as $Y = \{y_1, y_2, \cdots, y_m\}$ where n, m are the length of the X and Y, respectively[2].

3.2 Task

NMT is an approach to solve the machine translation problem. Different from traditional statistical machine translation which uses statistical methods, NMT uses artificial neural networks(ANN) to model the sentences and predict the translation results. With the development of deep neural networks, DNN is widely studied for NMT which can provide more precise outcomes.

3.3 Models

Here we just simply introduce the typical models for NMT including encoder-decoder model, RNN aligned model and Attention based models.

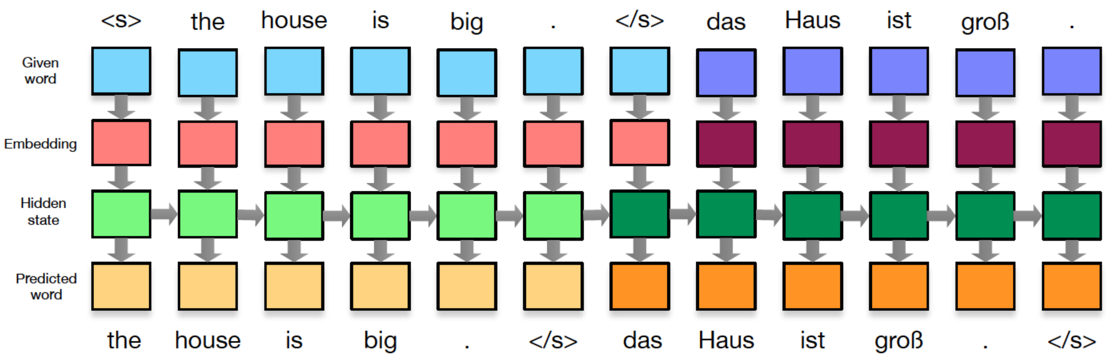

Figure 1: Sequence-to-sequence encoder-decoder model

3.3.1 Encoder-Decoder

The first introduced model is encoder-decoder model. Such a model can be decomposed as *encoder* and *decoder*, like the sequence-to-sequence model[7] referred as Figure 1.

The task of the encoder is to provide a representation of source sentence or text. Then the representation of the decoded source will go through the predictions of the model which output an end of sentence token. During this process, the hidden state of the source actually encodes the meanings and the vector which hold the values before going through decoder is called the *input sentence embedding*. After the *encoder phase*, the hidden state will be decoded in the *decoder phase*.

In decoder phase, the decoder should be able to predict the next word according to the information it can decode and also need to care about the coverage of the translation. So, the tasks of decoder cover two parts: prediction and translation.

3.3.2 RNN Aligned Model

Based on the previous encoder-decoder model, recurrent neural network(RNN) is aligned into the hidden state[4]. As the Figure 2, the left word will be passed to right in the first RNN and to get the right context a right-to-left RNN is added. Such a model is called *bidirectional recurrent neural network*. Denote the left-to-right hidden state as $\overrightarrow{h}$ and right-to-left as $\overleftarrow{h}$, then the mathematical representation of these two RNN are:

$$\begin{aligned}\overrightarrow{h_j} &= f(\overrightarrow{h_{j+1}}, \bar{E}x_j)\\ \overleftarrow{h_j} &= f(\overleftarrow{h_{j+1}}, \bar{E}x_j)\end{aligned}$$

where the x_j denotes an input word and for the general mapping function f, it can be feed-forward neural network, more complex recurrent unit(GRUs) or Long short-term memory(LSTM).

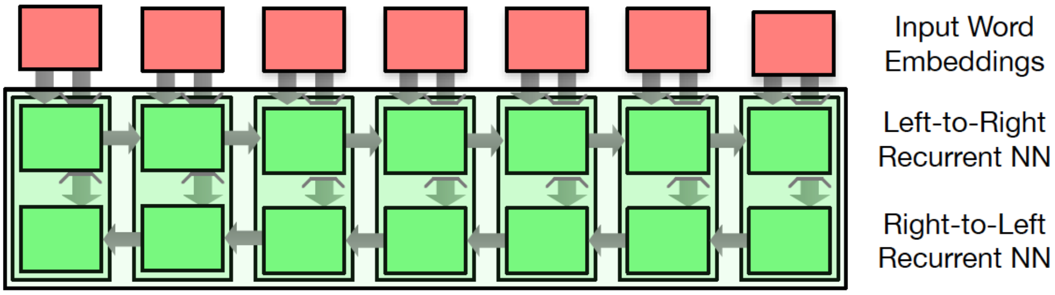

Figure 2: Input Encoder

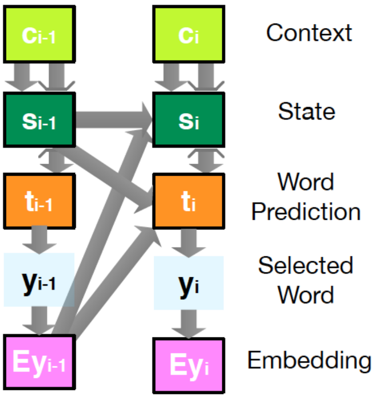

Figure 3: Output Decoder

Combined with hidden states, the decoder is also RNN(Figure 3) which takes the representation of the input context, the hidden states, and the output of the prediction to generate new hidden decoder state as well as word predictions.

The decoder will compute a sequence of hidden states s_i from previous hidden state s_{i-1}, the embedding of the previous output word Ey_{i-1}, and the input source c_i, and it is defined as:

$$s_i = f(s_{i-1}, Ey_{i-1}, c_i)$$

The function here also has multiple choices including GRUs, and LSTM, and so on.

In Figure 3, t_i is a conditioned values on the decoder hidden state s_{i-1}. So, the embedding of the previous output word Ey_{i-1} and the input text c_i can be represented as

$$t_i = \text{softmax}(W(U_{s_{i-1}} + VEy_{i-1} + Cc_i))$$

The softmax here can ensure the sum of the vector of the prediction probability be 1 and

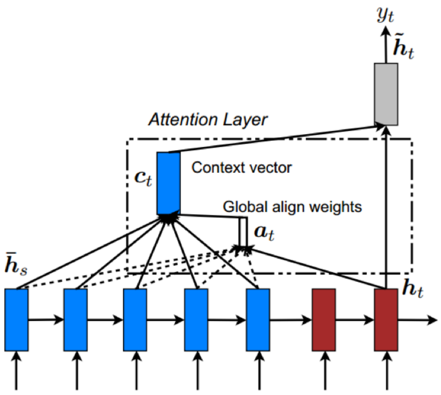

Figure 4: Global Attention Model

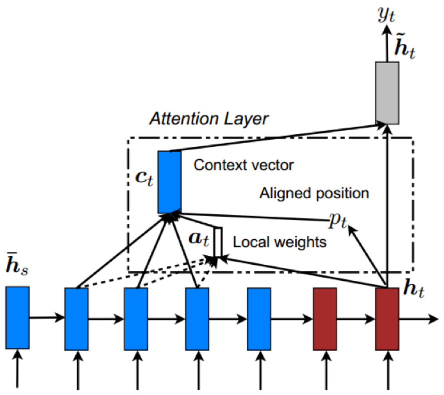

Figure 5: Local Attention Model

the max value indicates the output y_i.

3.3.3 Attention Based Model

Attention is a mechanism of NMT, it enables NMT to use the most relevant parts of the source sentence at each translation step. Attention is different from the alignment in some cases and is capturing useful information other than alignments[1]. This ability makes attentional NMT models perform better in handling long sentences.

There are two popular attention models: *global attention model* and *local attention model*. Common of these two models is in each time step t, attention layer uses the hidden state of LSTM $\tilde{h}_t$ as input to derive a context vector. The difference of the two models is the way they compute the context vector. Figure 4 and Figure 5[5] shows the difference.

In Figure 4, the global model infers a variable-length alignment weight vector at based on the current target state. In Figure 5, the local model first predicts a single aligned position for the current target word. A window centred around the source position is then used to compute a context vector. In global attention model, the attention layer focuses on the context of each word and the context vector contains the context information so that the model can use this information to perform better especially in translating long sentences.But if the sentence is much too long, the cost of calculation will be unacceptable. Thus, the local attention model is needed. Local attention model calculates the aligned position of each word first which gives the bound of the context of each word making the calculation of context vector easier.

3.4 Techniques

3.4.1 Beam Search

Beam search[6] is often used in statistical machine translation or neural machine translation to optimize the decoding process. In Figure 6, assume that the beam length is 2, that is, only 2 search paths are reserved at a time, and the decoder needs to output 2 target language words with the highest probability as candidates each time. And the decoder finally select the target language sequence with the highest probability as output.

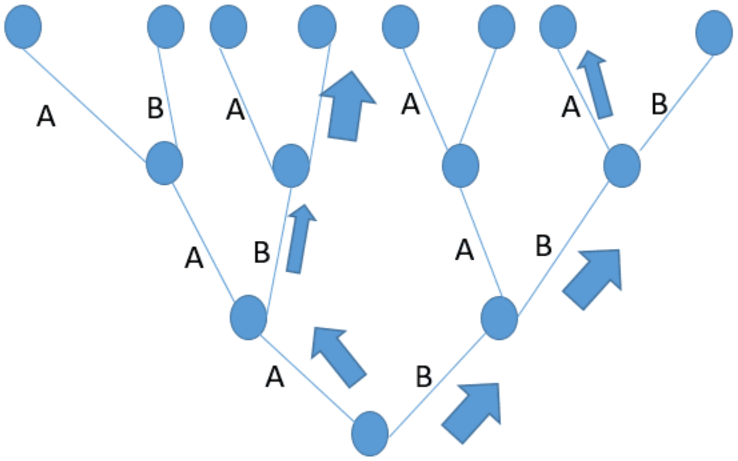

Figure 6: Beam Search(Length 2)

3.4.2 Residual Connection

Residual connection(Figure 7) is a cross-level connection mechanism[10]. Adding an identity mapping to the original RNN network, the original output is integrated with the direct input as the output, slowing down the rate of gradient descent to achieve the purpose of deepening the number of layers in the RNN network and reducing the difficulty of training.

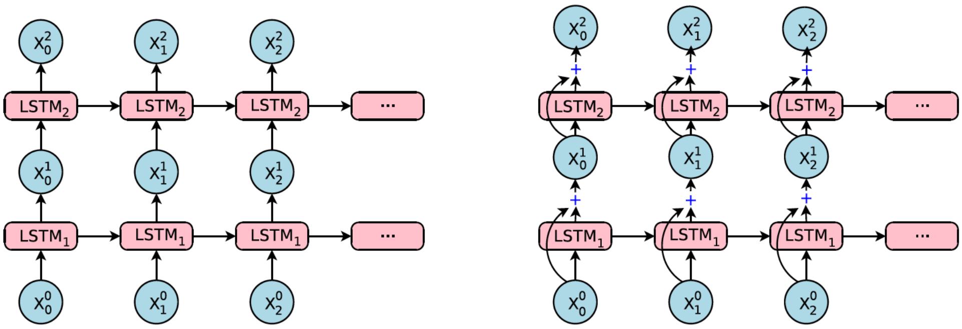

Figure 7: The difference of LSTM and Google's LSTM with residual connection

3.5 Example: Google's NMTS

In [10], Google's Neural Machine Translation System is a model consists of a deep LSTM with 8 encoder and 8 decoder layers using residual connection(Figure 7) as well as attention connections from the decoder network to the encoder. As Figure 8 shows, this model can be divided into three parts: encoder, decoder and the attention work. The encoder will transform the source sentence into a list of vectors in which every vector is a input symbol. The decoder will produces one symbol each time till the special end-of-sentence(EOS) is produced. The attention network connects the encoder and the decoder and make the decoder can focus on various regions of the source.

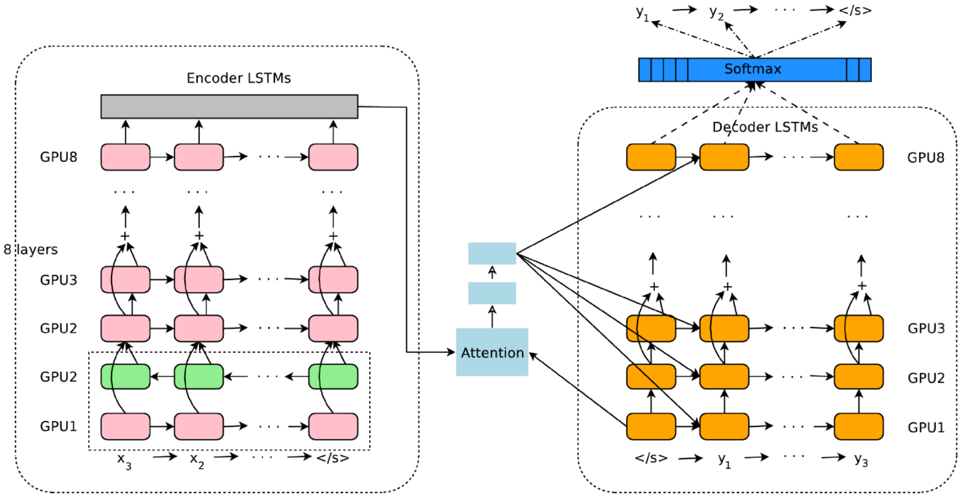

Figure 8: Google's Neural Machine Translation System

3.6 Datasets

Workshop on Statistical Machine Translation(WMT) is a widely used dataset post by Annual Meeting of the Association for Computational Linguistics. It contains training data, different tasks, and machine translation models. For example, in WMT 2018 eight language pairs are provided, along with a common framework. Its task is to improve current methods. Europarl is another widely used data set. It contains the data of parallel texts which are often from multinational institutions such as the United Nations, the European Union and the governments of multilingual countries. Europarl comprises of about 30 million words for each of the 11 official European languages.

- WMT Dataset

  The provided data is mainly taken from public data sources such as the Europarl corpus, and the UN corpus. Additional training data is taken from the News Commentary corpus.It allows participants to submit translations for any languages the data set contains.Human judges will judge the performance.The data has been divided into several tasks,including neural MT training task.This task requires the participants to train a fixed neural MT model with fixed data.Participants are required to submit the variables file i.e. the neural network.

  `http://www.statmt.org`

- Europarl Dataset[3]

  The Europarl parallel corpus is extracted from the proceedings of the European Parliament. It includes versions in 21 European languages: Romanic (French, Italian, Spanish, Portuguese, Romanian), Germanic (English, Dutch, German, Danish, Swedish), Slavik (Bulgarian, Czech, Polish, Slovak, Slovene), Finni-Ugric (Finnish, Hungarian, Estonian), Baltic (Latvian, Lithuanian), and Greek.

  `http://www.statmt.org/europarl/`

- Linguistic Data Consortium(LDC)

  The Linguistic Data Consortium (LDC) is an open consortium of universities, libraries, corporations and government research laboratories. LDC initially was a repository and distribution point for language resources. With the development of the LDC, it then changed to be an organization and nowadays it supports language resources for language-based technology evaluations.

  `https://www.ldc.upenn.edu`

- OPUS[9]

  OPUS is a growing collection of translated texts from the web. The main Chinese dataset is *MultiUN* and *OpenSubtitles2016*.

`http://opus.lingfil.uu.se/MultiUN.php`

`http://opus.lingfil.uu.se/OpenSubtitles2016.php`

`http://opus.lingfil.uu.se/`

- Acquis Communautaire(AC)

  As there were 20 official EU languages at the beginning of the year 2005, the AC thus exists as a parallel text (text and its translation) in 20 languages. The languages are Czech, Danish, German, Greek, English, Spanish, Estonian, Finnish, French, Hungarian, Italian, Lithuanian, Latvian, Maltese, Dutch, Polish, Portuguese, Slovak, Slovene and Swedish.

  `https://wt-public.emm4u.eu/Acquis/JRC-Acquis.2.2/doc/README_Acquis-Communautaire-corpus_JRC.html`

- UM-Corpus[8]

  The UM-Corpus has been designed to be a multi-domain and balanced parallel corpus for research and development purpose. In this version, a two million English-Chinese aligned corpus is provided, and it is categorized into eight different text domains, covering several topics and text genres, including: Education, Laws, Microblog, News, Science, Spoken, Subtitles, and Thesis.

  `http://nlp2ct.cis.umac.mo/um-corpus/index.html`

3.7 Evaluation

3.7.1 Bilingual Evaluation Understudy(BLEU)

The essence of *BLEU* is the calculation of the co-occurrence frequency of two sentences[6]. BLEU's design philosophy is consistent with the idea of judging good or bad machine translation: The closer the machine translation results are to the results of professional human translation, the better. What the BLEU algorithm is actually doing: judging the similarity of two sentences. In the BLEU, If a machine translation is compared with its corresponding reference translation, a comprehensive score is calculated. The higher the score, the better the machine translates.

The process to compute the BLEU is that firstly compute p_n, where p_n is a modified precision score named modified n-gram precision. The molecular is the sum of the clipped n-gram counts for all the candidate sentences which appear in references and the denominator is the number of candidate n-grams in the test corpus.

$$p_n = \frac{\sum_{\mathcal{C}\in\{Candidates\}} \sum_{n-gram\in\mathcal{C}} Count_{clip}(n-gram)}{\sum_{\mathcal{C}'\in\{Candidates\}} \sum_{n-gram'\in\mathcal{C}'} Count_{clip}(n-gram')}$$

Let c be the length of the candidate translation and r be the effective reference corpus length. It compares the length of the candidates and the references. Compute the brevity penalty BP:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{1}$$

BP can resolve the problem of judging whether the translation is complete.

Then,

$$BLEU = BP \cdot \left(\sum_{n=1}^{N} w_n \log p_n \right)$$

BLEU value is calculated by combining the p_n values corresponding to different lengths n.

If use the log domain, the ranking behaviour will more immediately apparent:

$$\log BLEU = \min{(1 - \frac{r}{c}, 0)} + \sum_{n=1}^{N} w_n \log p_n$$

3.8 Summary

This part is a simplified survey including the basic model, a widely used technique and datasets of NMT. The model is very simple to understand in structure. RNN, LSTM are normally used to construct a deep neural network translation system. The attention mechanism is an important technique in NMT, two attention based models are introduced. For training ANN, the dataset is crucial and the well-known datasets are introduced. An evaluation BLEU is introduced since the measurement of model performance is a key point.

4 Research Approach

4.1 Goal

The goal of this project is an Android application which provides NMT through an off-line neural network. The application can accept users' voice and output corresponding translation results. The basic requirement of the translation service should include but not limit to Chinese-to-English and English-to-Chinese.

4.2 Steps

The process can be divided into steps:

1. Target deep neural network models for NMT
2. Adapt DNN model and deploy into application
3. Test and optimize application on real mobile devices

4.3 Timeline

The timeline of this this project can be referred as Figure 9.

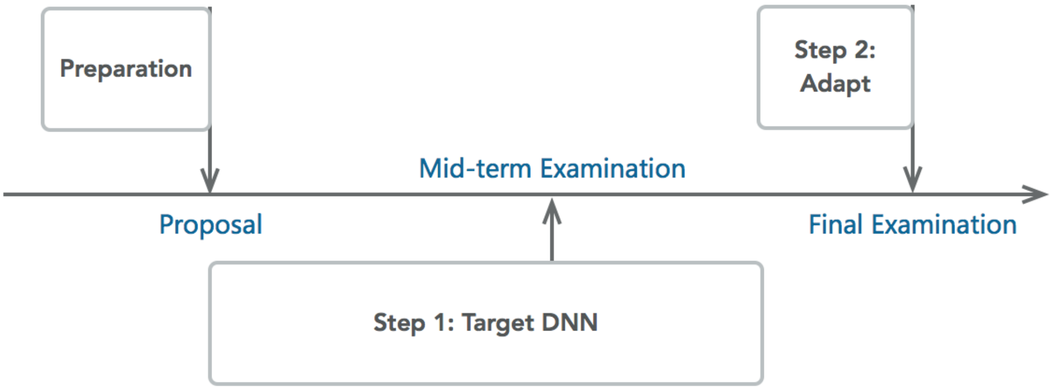

Figure 9: Timeline

5 Technology Roadmap

The rough roadmap of this project can be represent as Figure 10

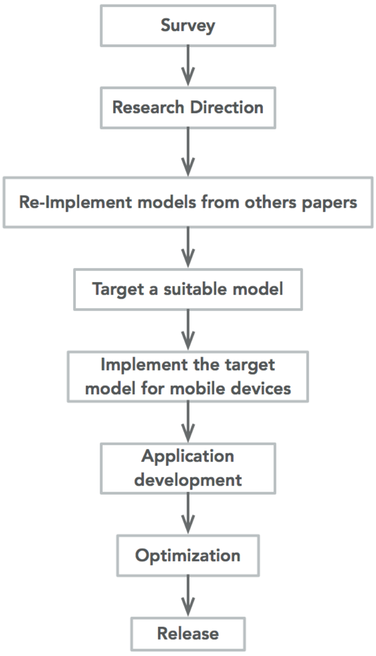

Figure 10: Technology Roadmap

6 Experiment Settings

The first part is the core of this project is now our focus. After our primary research, we are going to reproduce the models in some papers we have read. Multiple datasets will be used to test various models and the best-performed model will be selected as the target model. After that, neural network compression will be the new challenge of our project, we will then focus on compressing our DNN model for mobile devices. The final model will be a compressed deep neural network model integrated into an Android application.

7 Conclusion

So far, we have already finished our primary research and have comprehended the basic model and method of NMT. NMT focuses on handling machine translation problem by using sequence handling ability of deep neural networks. We also have some understanding of models and techniques of NMT as well as the data set and its evaluation standard. In the future, we are going to choose one or more models and algorithms in the papers to reproduce. After then we will focus on Neural network compression to deploy the models on mobile devices. At last, we will combine our product with voice technology, so that our problem can be solved.

References

[1] Hamidreza Ghader and Christof Monz. What does attention in neural machine translation pay attention to? *arXiv preprint arXiv:1710.03348*, 2017.

[2] Di He, Hanqing Lu, Yingce Xia, Tao Qin, Liwei Wang, and Tieyan Liu. Decoding with value networks for neural machine translation. In *Advances in Neural Information Processing Systems*, pages 177–186, 2017.

[3] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

[4] Philipp Koehn. Neural machine translation. *arXiv preprint arXiv:1709.07809*, 2017.

[5] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

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

[8] Liang Tian, Derek F Wong, Lidia S Chao, Paulo Quaresma, Francisco Oliveira, and Lu Yi. Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In *LREC*, pages 1837–1842, 2014.

[9] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[10] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.