EFFICIENT MODEL FOR ENGLISH TO SERBIAN NEURAL MACHINE TRANSLATION

Dragutin Ostojić[1], Branko Arsić[1], Tatjana Stojanović[1], Neda Vidanović Miletić[2]

[1] University of Kragujevac, Faculty of Science, 12 Radoja Domanovića Street, 34000 Kragujevac
Email: dragutin.ostojic@pmf.kg.ac.rs, brankoarsic@kg.ac.rs, tanjat@kg.ac.rs
[2] University of Kragujevac, Faculty of Engineering, 6 Sestre Janjić Street, 34000 Kragujevac
Email: neda@kg.ac.rs

Abstract

Machine Translation (MT) refers to the creation of an automatic system for translation from one language into another. Although the idea of MT is not novel, it has become a sustainable tool in widespread use in the past ten years. The deep learning-based approach to MT, neural MT (NMT), has progressed rapidly in recent years, but is barely applied to South Slavic languages. In this paper, we have studied the specific case of translation from English to Serbian, focusing on adjustments of parameters on existing artificial neural network (ANN) architectures for efficient training. We also performed careful data preparation using custom methods. The result is a NMT model whose performances provide the opportunity for training even on a PC, while the effectiveness of the translations themselves does not lag far behind the industry leaders in the field.

Key words: Machine translation, Performance trade-off, Serbian to English

1 Introduction

MT today refers to the problem of creating software for automatic translation. It has been studied the most on Russian and English, later adding French, Japanese, Chinese and other languages [2]. However, South Slavic languages are barely analysed. The reasons can be found in their complexity because of rich morphology, complex grammar, several dialects, two alphabets, etc., which old methods cannot handle properly [4].

Traditional approaches like rule-based machine translation (RBMT) and example-based machine translation (EBMT) are problematic due to high expenses, great human input, or a lack of good examples. Also, the newer method statistical machine translation (SMT) needs massive parallel corpora and it is hard to correct inaccuracies [2].

NMT is the newest approach based on recurrent ANNs (RNN) ability to extract properties from data sequences on their own. Only in recent years has this technology, as well hardware possibilities, evolved enough to be used in practice. It has already proved to be successful, with great potential for further improvements as well [1, 3].

This paper describes the process of creating an NMT model for English to Serbian translation. The development incorporates the latest achievements in this field, emphasizing the retention of model simplicity and the possibility of training it in a reasonable timeframe even on a PC, without sacrificing the quality of translation.

We emphasized the delicate matter of data choice and preparation. The goal of this paper is the consideration of possibilities and the requirements of currently available technologies, as well as the establishment of the basis for further research.

2 Methodology

We constructed an encoder-decoder experimental model (EM) (Fig. 1). Model input is a sentence of One-hot encoded words with symbols for the start and the end of a sequence. The embedding layer transforms it to fixed-size vectors based on their meaning. The encoder extracts information of the sentence into a fixed-size Context vector with bidirectional Long short-term memory (LSTM), a kind of RNN, layers. Finally, the decoder is constructing translation based on LSTM layers, using the Attention mechanism. For the efficiency process of training, we used the Teacher forcing technique [7, 8].

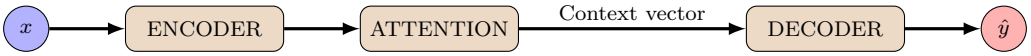

Figure 1: Encoder-Decoder architecture with attention mechanism

The BLEU score [6] was used as the main metric for measuring results. We also used NLLLoss and Perplexity for monitoring the training process. Next, we implemented our data preparation techniques, and tried numerous hyperparameter values and found the best ones. Afterwards, we compared our test results with popular translators.

3 Data preparation

Publicly available English-Serbian parallel corpora that have 10^4 translation pairs at least, and correct translation quality in random sampling are: *Tatoeba* - translated sentences, *SETIMES* - newspaper texts, and *QED* - educational videos subtitles [10].

The rough preparation of corpora is performed in four phases which refer to: alphabet reduction and unification, syntax simplification, duplicate elimination, and lexicon filtering. Afterwards, NMT preparation filters bad pairs using some pre-trained model.

Each prepared corpus is divided into a training and test set in ratio 9:1.

4 Experimental setup

Using enormous resources owned by *Google*, a series of different NMT models were tested and a very important ascertainment was made that, in the case of MT, more does not necessarily mean better [9]. The example is the discovery that the increase of the neural network depth, after a certain point, can lead not only to the increase of the time needed for training, but also to a worse result. Based on other studies [5, 7, 8, 9], as well as our numerous experiments related to this specific case, we delineated the design and set the values for the hyperparameters of the EM, as seen in Table 1.

Hyperparameter	Value	Hyperparameter	Value
RNN cells type	LSTM	Input encoding	One-hot
Embedding vector size	512	Loss function	NLLLoss
Context vector size	512	Batch size	128
Encoder type	Bidirectional	Dropout	0.2
Encoder depth	2	Optimizer	Adam
Decoder depth	2	Learning rate	0.0002
Attention type	Global	Stopping criteria	16 epochs BLEU score stagnation

Table 1: Model hyperparameters.

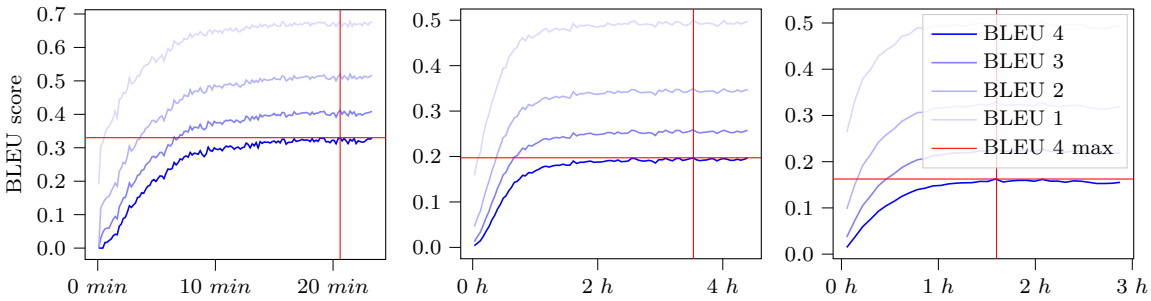

Figure 2: EM training on rough prepared test corpora *Tatoeba*, *SETIMES* and *QED*

	Tatoeba	SETIMES	QED
Google Translate	35.26	33.12	27.11
Microsoft Translator	32.65	52.27	24.00
Yandex Translate	20.58	24.62	19.26
EM	33.01	19.72	16.25

Table 2: BLEU 4 scores comparison on rough prepared test corpora

Fig. 2 shows a comparison of model performances for test corpora. Table 2 represents the comparison of EM and industrial leaders' models on given test corpora.

In the next experiment, the union of roughly prepared corpora is used for the training of EM that will be used for NMT data preparation. In this phase, the given paired translation is compared to the translation done by the trained model. If the BLEU score of the two translations exceeds 0.2, the pair is declared valid and kept in the corpus. This method is very significant as it efficiently enough removes the wrongly paired translation pairs common in automatically generated corpora, like *OpenSubtitles* [10]. Without this phase, those corpora would have no use in MT. Moreover, we proved that this method increases the effectiveness of MTs in general.

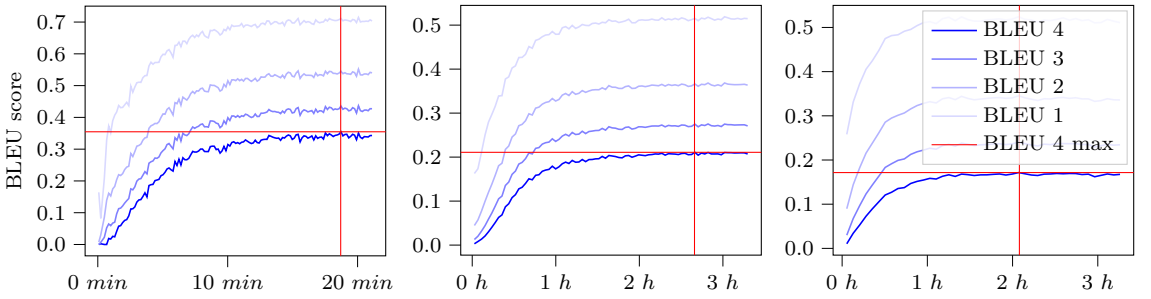

Figure 3: EM training on rough and NMT prepared test corpora *Tatoeba*, *SETIMES* and *QED*

	Tatoeba	SETIMES	QED
Google Translate	38.15	38.49	30.91
Microsoft Translator	33.82	55.38	26.03
Yandex Translate	20.52	26.11	20.98
EM	35.47	21.09	17.16

Table 3: BLEU 4 scores comparison on rough and NMT prepared test corpora

The results after the second phase are shown in Fig. 3. Table 3 represents the comparison of EM and industrial leaders' models on given test corpora.

All the tests were performed on *GIGABYTE NVIDIA GeForce GTX 1080 Ti 12GB GDDR5X 352bit (driver 440.82 and CUDA driver 10.2), Intel Core i5-6400 CPU @ 2.70GHz and 8GB RAM* with *PyTorch 1.9, Python 3.8, Ubuntu 18.04*.

5 Conclusions

This paper shows an example of the use of the NMT approach for English-Serbian translation, as well as a comparison of the performances of the obtained model with known commercial MTs. The implementation included various techniques for the improvement of the translation quality and performances related to the very process of training. The main advantages of the constructed model are the simplicity and low hardware complexity that leave room for further improvement. However, the main disadvantage of the constructed model is the generally lower translation quality compared to more famous competitors. The main reason for this is the lack of adequate language corpora and the complexity of the Serbian language. Future research will be focusing on the model improvement, its training on larger corpora and translating from Serbian to English too, as well as the creation of methods for new corpus generation and testing new approaches to MT, like the Transformer model [11].

Acknowledgements: This paper has been supported by the Serbian Ministry of Education, Science and Technological Development, Grants No. 451-03-68/2022-14/ 200122. This research has also been funded through the EIT's HEI Initiative SMART-2M project, supported by EIT RawMaterials, funded by the European Union.

References

[1] D. Jurafsky, J.H. Martin, Speech and Language Processing, Prentice Hall, 2009.

[2] W.J. Hutchins, Early Years in Machine Translation, JB Publishing Company, 2000.

[3] K. Cho et al., Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. EMNLP, 2014.

[4] P. Lohar et al., Building English-to-Serbian Machine Translation System for IMDb Movie Reviews, BSNLP, 2019.

[5] D. Britz et al., Massive exploration of neural machine translation architectures, EMNLP, 2017.

[6] K. Papineni et al., BLEU, ACL, 2002.

[7] S. Robertson, NLP From Scratch, PyTorch, 2018.

[8] Q. M. Lanners, T. Laurent, Neural Machine Translation, 2019.

[9] Y. Wu et al., Google's neural machine translation system, arXiv, 2016.

[10] J. Tiedemann, Parallel Data, Tools and Interfaces in OPUS, LREC, 2012.

[11] Ashish Vaswani et al., Attention Is All You Need, NIPS, 2017.