JUMT at WMT2019 News Translation Task: A Hybrid approach to Machine Translation for Lithuanian to English

Sainik Kumar Mahata, Avishek Garain, Adityar Rayala,
Dipankar Das, Sivaji Bandyopadhyay
Computer Science and Engineering
Jadavpur University, Kolkata, India
sainik.mahata@gmail.com, avishekgarain@gmail.com, mailsofadityar@gmail.com,
dipankar.dipnil2005@gmail.com, sivaji_cse_ju@yahoo.com

Abstract

In the current work, we present a description of the system submitted to WMT 2019 News Translation Shared task. The system was created to translate news text from Lithuanian to English. To accomplish the given task, our system used a Word Embedding based Neural Machine Translation model to post edit the outputs generated by a Statistical Machine Translation model. The current paper documents the architecture of our model, descriptions of the various modules and the results produced using the same. Our system garnered a BLEU score of 17.6.

1 Introduction

Machine **T**ranslation (MT) is automated translation of one natural language to another using a computer. Translation, itself, is a very tough task for both humans as well as a computer. It requires a thorough understanding of the syntax and semantics of both the languages under consideration. For producing good translations, a MT system needs good quality and sufficient amount of parallel corpus (Mahata et al., 2016, 2017).

In the modern context, MT systems can be categorized into **S**tatistical **M**achine **T**ranslation (SMT) and **N**eural **M**achine **T**ranslation (NMT). SMT has had its share in making MT very popular among the masses. It includes creating statistical models, whose input parameters are derived from the analysis of bilingual text corpora, created by professional translators (Weaver, 1955). The state-of-art for SMT is Moses Toolkit[1], created by Koehn et al. (2007), incorporates subcomponents like Language Model generation, Word Alignment and Phrase Table generation. Various works have been done in SMT (Lopez, 2008; Koehn, 2009) and it has shown good results for many language pairs.

On the other hand NMT (Bahdanau et al., 2014), though relatively new, has shown considerable improvements in the translation results when compared to SMT (Mahata et al., 2018b). This includes better fluency of the output and better handling of the Out-of-Vocabulary problem. Unlike SMT, it doesnt depend on alignment and phrasal unit translations (Kalchbrenner and Blunsom, 2013). On the contrary, it uses an Encoder-Decoder approach incorporating Recurrent Neural Cells (Cho et al., 2014). As a result, when given sufficient amount of training data, it gives much more accurate results when compared to SMT (Doherty et al., 2010; Vaswani et al., 2013; Liu et al., 2014).

For the given task[2], we attempted to create a MT system that can translate sentences from Lithuanian to English. Since, using only SMT or NMT models leads to some or the other disadvantages, we tried to use both in a pipeline. This leads to an improvement of the results over the individual usage of either SMT or NMT. The main idea was to train a SMT model for translating Lithuanian language to English. Thereafter, a test set was translated using this model. Then, a word embedding based NMT model was trained to learn the mappings between the SMT output (in English) and the gold standard data (in English).

The organizers provided the required parallel corpora, consisting of 9,62,022 sentence pairs, for training the translation model. Among this, 7,62,022 pairs was used to train the SMT system and 2,00,000 pairs were used to test the SMT system and then train the NMT system. The statistics of the parallel corpus is depicted in 1.

The remainder of the paper is organized as follows. Section 2 will describe the methodology of creating the SMT and the NMT model and will in-

[1] http://www.statmt.org/moses/

[2] http://www.statmt.org/wmt19/translation-task.html

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 283–286
Florence, Italy, August 1-2, 2019.

# sentences in Lt corpus	9,62,022
# sentences in En corpus	9,62,022
# words in Lt corpus	1,16,65,937
# words in En corpus	1,56,22,488
# word vocab size for Lt corpus	4,88,593
# word vocab size for En corpus	2,27,131

Table 1: Statistics of the Lithuanian-English parallel corpus provided by the organizers. "#" depicts No. of. "Lt" and "En" depict Lithuanian and English, respectively. "vocab" means vocabulary of unique tokens.

clude the preprocessing steps, a brief summary of the encoder-decoder approach and the architecture of our system. This will be followed by the results and conclusion in Section 3 and 4, respectively.

2 Methodology

2.1 SMT

For designing the model we followed some standard preprocessing steps on 7,62,022 sentence pairs, which are discussed below.

2.1.1 Preprocessing

The following steps were applied to preprocess and clean the data before using it for training our Statistical machine translation model. We used the NLTK toolkit[3] for performing the steps.

- **Tokenization**: Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens. In our case, these tokens were words, punctuation marks, numbers. NLTK supports tokenization of Lithuanian as well as English texts.
- **Truecasing**: This refers to the process of restoring case information to badly-cased or non-cased text (Lita et al., 2003). Truecasing helps in reducing data sparsity.
- **Cleaning**: Long sentences (of tokens > 80) were removed.

2.1.2 Moses

Moses is a statistical machine translation system that allows you to automatically train translation models for any language pair, when trained with a large collection of translated texts (parallel corpus). Once the model has been trained, an efficient search algorithm quickly finds the highest probability translation among the exponential number of choices.

We trained Moses using 7,62,022 sentence pairs provided by WMT2019, with Lithuanian as the source language and English as the target language. For building the Language Model we used KenLM[4] (Heafield, 2011) with 7-grams from the target corpus. The English monolingual corpus from WMT2019 was used to build the language model

Training the Moses statistical MT system resulted in generation of Phrase Model and Translation Model that helps in translating between source-target language pairs. Moses scores the phrase in the phrase table with respect to a given source sentence and produces best scored phrases as output.

2.2 NMT

Neural machine translation (NMT) is an approach to machine translation that uses neural networks to predict the likelihood of a sequence of words. The main functionality of NMT is based on the sequence to sequence (seq2seq) architecture, which is described in Section 2.2.1.

2.2.1 Sequence to Sequence Model

Sequence to Sequence learning is a concept in neural networks, that helps it to learn sequences. Essentially, it takes as input a sequence of tokens (words in our case)

$$X = \{x_1, x_2, ..., x_n\}$$

and tries to generate the target sequence as output

$$Y = \{y_1, y_2, ..., y_m\}$$

where x_i and y_i are the input and target symbols respectively.

Sequence to Sequence architecture consists of two parts, an Encoder and a Decoder.

The encoder takes a variable length sequence as input and encodes it into a fixed length vector, which is supposed to summarize its meaning and taking into account its context as well. A **L**ong **S**hort **T**erm **M**emory (LSTM) cell was used to achieve this. The uni-directional encoder reads the words of the Lithuanian texts, as a sequence from one end to the other (left to right in our case),

$$\vec{h}_t = \vec{f}_{enc}(E_x(x_t), \vec{h}_{t-1})$$

[3]https://www.nltk.org/

[4]https://kheafield.com/code/kenlm/

Here, E_x is the input embedding lookup table (dictionary), $\vec{f}_{enc}$ is the transfer function for the LSTM recurrent unit. The cell state *h* and context vector *C* is constructed and is passed on to the decoder.

The decoder takes as input, the context vector *C* and the cell state *h* from the encoder, and computes the hidden state at time t as,

$$s_t = f_{dec}(E_y(y_{t-1}), s_{t-1}, c_t)$$

Subsequently, a parametric function out_k returns the conditional probability using the next target symbol k.

$$(y_t = k \mid y < t, X) = \frac{1}{Z} exp(out_k(E_y(y_t - 1), s_t, c_t))$$

Z is the normalizing constant,

$$\sum\nolimits_j exp(out_j(E_y(y_t - 1), s_t, c_t))$$

The entire model can be trained end-to-end by minimizing the log likelihood which is defined as

$$L = -\frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_y{}^n} logp(y_t = y_t{}^n, y_{it}{}^n, X^n)$$

where N is the number of sentence pairs, and X^n and $y_t{}^n$ are the input sentence and the t-th target symbol in the n-th pair respectively.

The input to the decoder was one hot tensor (embeddings at word level) of 2,00,000 English sentences while the target data was identical, but with an offset of one time-step ahead.

2.3 Architecture

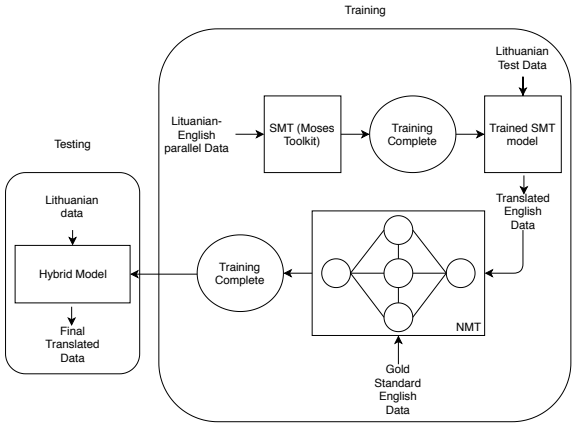

Figure 1: Architecture

2.3.1 Training

For the training purpose, 7,62,202 , preprocessed, Lituanian-English sentence pairs were fed to Moses Toolkit. This created a SMT translation model with Lithuanian as the source language and English as the target language. Thereafter, we had 2,00,000 Lithuanian-English sentence pairs, from which the Lithuanian sentences were given as input to the SMT model and it gave 2,00,000 translated English sentences as output. Now, this 2,00,000 translated English sentences and the respective gold standard 2,00,000 sentences, from the Lithuanian-English sentence pair, were given as input to a word embedding based NMT model. As a result, this constituted our Hybrid model.

2.3.2 Testing

For the testing purpose, 10k Lithuanian Sentences were fed to the Hybrid model, and the output, when checked using BLEU (Papineni et al., 2002), resulted in an accuracy of 21.6. The training and testing architecture is shown in Figure 1

3 Results

WMT2019 provided us with a test set of Lithuanian sentences in .SGM format. This file was parsed and fed to our hybrid system. The output file was again converted to .SGM format and submitted to the organizers. Our system garnered a BLEU Score of 17.6, when it was scored using automated accuracy metrics. Other accuracy scores are mentioned in Table 2.

Metric	Score
BLEU	17.6
BLEU-cased	16.6
TER	0.762
BEER 2.0	0.497
CharactTER	0.718

Table 2: Accuracy scores calculated using various autmoated evaluation metrics.

4 Conclusion

The paper presents the working of the translation system submitted to WMT 2019 News Translation shared task. We have used Word Embedding based NMT on top of SMT, for our proposed system. We have used a single LSTM layer as an encoder as well as a decoder. As a future prospect, we plan to use more LSTM layers in our model. We plan

to create another model that incrementally trains both the SMT and NMT systems in a pipeline to improve the translation quality.

Acknowledgement

The reported work is supported by Media Lab Asia, MeitY, Government of India, under the Visvesvaraya PhD Scheme for Electronics & IT.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Stephen Doherty, Sharon O?Brien, and Michael Carl. 2010. Eye tracking as an mt evaluation technique. *Machine translation*, 24(1):1–13.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *EMNLP*, volume 3, page 413.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 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*, pages 177–180. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *CoRR*, abs/1610.03017.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W. Black. 2015. Character-based neural machine translation. *CoRR*, abs/1511.04586.

Lucian Vlad Lita, Abe Ittycheriah, Salim Roukos, and Nanda Kambhatla. 2003. Truecasing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 152–159. Association for Computational Linguistics.

Shujie Liu, Nan Yang, Mu Li, and Ming Zhou. 2014. A recursive recurrent neural network for statistical machine translation.

Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):8.

Sainik Mahata, Dipankar Das, and Santanu Pal. 2016. Wmt2016: A hybrid approach to bilingual document alignment. In *WMT*, pages 724–727.

Sainik Kumar Mahata, Dipankar Das, and Sivaji Bandyopadhyay. 2017. Bucc2017: A hybrid approach for identifying parallel sentences in comparable corpora. *ACL 2017*, page 56.

Sainik Kumar Mahata, Dipankar Das, and Sivaji Bandyopadhyay. 2018a. Jucbnmt at wmt2018 news translation task: Character based neural machine translation of finnish to english. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 445–448.

Sainik Kumar Mahata, Dipankar Das, and Sivaji Bandyopadhyay. 2018b. Mtil2017: Machine translation using recurrent neural network on statistical machine translation. *Journal of Intelligent Systems*, pages 1–7.

Sainik Kumar Mahata, Soumil Mandal, Dipankar Das, and Sivaji Bandyopadhyay. 2018c. Smt vs nmt: A comparison over hindi & bengali simple sentences. *arXiv preprint arXiv:1812.04898*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. Decoding with large-scale neural language models improves translation. In *EMNLP*, pages 1387–1392.

Warren Weaver. 1955. Translation. *Machine translation of languages*, 14:15–23.