University *of Ljubljana*
Faculty *of Computer and Information Science*

Neural machine translation

Erika Stanković, Eva Boneš, Sara Sever, Meta Jazbinšek, Teja Hadalin

Abstract

task selection simple corpus processing/analysis Introduction, existing solutions, initial ideas

Keywords
Keyword1, Keyword2, Keyword3 ...

Advisors: Slavko Žitnik, Špela Vintar, Mojca Brglez

1. Introduction

Machine translation (MT) is the problem of automatically translating text from one language to another. It has been researched since the 1950s, but only recently, with the rise of deep learning, did it prove to be solvable, although the possibility of achieving fully automatic machine translations of high quality is still being questioned.

There are three main approaches to solving the problem, all with their advantages and shortcomings. Rule-based machine translation (RBMT) is the oldest of the bunch and it requires expert knowledge about both the source and the target language to develop syntactic, semantic, and morphological rules. In contrast to the other two approaches, RBMT does not need bilingual data and when a set of rules is set for a particular language it can be reused for pairings with different languages. Its main disadvantage is that it requires a lot of expertise to manually set rules and good dictionaries which can take a lot of time or might even border on impossible in some languages.

Another approach, which gained popularity in the 1990s, uses statistical models based on the analysis of bilingual text corpora. The idea behind statistical machine translation (SMT) as proposed in [1] is, if given a sentence T in the target language, we seek the sentence S from which the translator produced T – we want to choose S so that it maximizes $Pr(S|T)$. Comparing to RBMT, this approach needs less manual work from linguistic experts and we can achieve more fluent translations (provided we have a good language model). STM models are also generally built in a way that one can be used on more language pairs, but this requires a bilingual corpus which can be hard to find, depending on the language pair we are interested in. This model is also less suitable for language pairs with big differences in word order.

As with many computer science fields, the current state-of-the-art approaches are based on neural networks. Contrary to building a pipeline of separate tasks with RBMT and STM, we only need to build one network with neural machine translation (NMT). The biggest challenge in building a successful English to Slovene (or vice-versa) automatic translator is obtaining a sufficiently big bilingual corpus. Like all deep learning approaches, having a large and quality dataset is crucial for the success of the model. To deal with this exact problem, a lot of approaches of pre-training a network on monolingual data (that can be much easily obtained) have been proposed.

Bidirectional Encoder Representations from Transformers (BERT) [2] which was proposed by researchers at Google, uses two strategies to deal with the problem, namely masked language modeling (MLM) and next sentence prediction (NSP). With MLM before feeding sequences into the model, 15% of the words are replaced with a [MASK] token, and the model then attempts to reproduce the original value of sequences. This way, along with a bigger dataset we can use this way, the model also achieves more context awareness and when fine-tuned produces more fluent translations. In NSP, the model receives pairs of sentences as input and learns to predict if the second sentence is a subsequent sentence of the first one. Similar to MLM, the model achieves more context-awareness.

In 2020, mRASP [3], that could be very useful in our case, has been proposed. The authors built a pretrained NMT model that could be fine-tuned for any language pair. They used 197M pairs of sentences, which is much more than we could obtain for only English-Slovene translations.

Although these methods have been proven to be successful, one of today's largest available databases of pretrained translation models was trained using just a standard transformer model and still achieves great results. The Tatoeba Translation Challenge [4] aims to provide data and tools for

creating state-of-the-art translation models. The focus is on low-resource languages to push their coverage and translation quality. It currently includes data for 2,963 language pairs covering 555 languages. Along with the data, they also released pretrained translation models for multiple languages which are regularly updated. We have decided to use the English-Slovene pretrained model for the baseline to compare our translational models to.

Alongside the pretrained model, we trained a transformer model from scratch on a large general corpus, which we then fine-tuned on a corpus consisting of TED talks to make a translational model specialized for speech transcription translation. We evaluated all the models on the same validation set, both manually and automatically.

In Section 2, we first describe the data we used. In the subsequent Section, we describe all the methods for both training and evaluating the models. Later on, in Sections 4 and 5, we present the results and discuss them.

1.1 Frameworks

fairseq [5]

fairseq is a sequence modeling toolkit written in PyTorch for training models for translation, summarization, and other natural language processing tasks. It provides different neural network architectures, namely convolutional neural networks (CNN), Long-Short-Term Memory (LSTM) networks, and Transformer (self-attention) networks. The architectures can be configured to our own needs and many implementations for different tasks have been proposed since the fairseqs introduction in 2019. In addition to different architectures, they also provide pre-trained models and pre-processed test sets for different tasks, but none of them is in Slovene, so we won't be able to use them.

Marian NMT [6]

Marian is an efficient Neural Machine Translation framework written in pure C++. It is developed by Microsoft and it provides fast multi-GPU training and batched translation on GPU/CPU. It allows training on raw texts using built-in SentencePiece [7] (an unsupervised text tokenizer and detokenizer). It offers multiple different models: deep RNNs with Deep Transition Cells [8], transformer models [9], multi-source models [10], RNN, and transformer-based language models. The code is publicly available on https://marian-nmt.github.io/.

T5 [11]

T5 (Text-To-Text Transfer Transformer) is a framework, developed by Google, that trains a single model for a variety of text tasks (translation, summarization, ...). Its performance is comparable to task-specific architectures and was at that time considered state-of-the-art when combined with scale. It uses an encoder-decoder model and it was developed using Tensorflow. For tokenizing, it also uses the SentencePiece model [7]. The code is publicly available on https://github.com/google-research/text-to-text-transfer-transformer. The framework includes some pre-trained models, but because of the vocabulary they used, it can only process a predetermined, fixed set of languages. This means that for our task of translating from and into Slovene we would need to train the model from scratch. The framework needs a significant amount of preprocessing because it requires a specific format of input data: <input>\t<target>.

2. Data

2.1 General translation model

The datasets for the general translation model are the eight biggest corpora from the Opus site (https://opus.nlpl.eu [12]) for the Slovene-English language pair. The exact size of each one, complete with the number of tokens, links, sentence pairs, and words, is noted in Table 1. The corpora were chosen based on the quantity of the data, so the general translation model would contain a large amount of diverse information. After a brief look at the contents of each one, we can see that some datasets are of higher quality and more reliable, because of the source of the original texts and their translations, for example, the corpora from European institutions.

Europarl is a parallel corpus from the European Parliament website by Philipp Koehn (University of Edinburgh) and is extracted from the proceedings of the European Parliament, from 1996–2011. The **DGT** corpus is a collection of translation memories from the European Commission's Directorate-General for Translation. The aligned translation units in the DGT have been extracted from one of its large shared translation memories in EURAMIS (European advanced multilingual information system).

Other corpora are a collection of translations from different Internet sources, which makes them less reliable, however, they are still important to us because of the quantity of the data.

The **CCAligned** corpus consists of parallel or comparable web-document pairs in 137 languages aligned with English. These web-document pairs were constructed by performing language identification on raw web documents, and ensuring corresponding language codes were corresponding in the URLs of web documents. The **MultiCCAligned v1** corpus has been further processed for making it a multi-parallel corpus by pivoting via English. The **OpenSubtitles** is a corpus compiled from a large database of movie and TV subtitles from the http://www.opensubtitles.org/ repository. Based on their website, there are 745.363 subtitles for English and 72.819 for the Slovene language currently uploaded. From a linguistic perspective, subtitles cover a wide and interesting width of genres, from colloquial language or slang to narrative and expository discourse (as in e.g. documentaries) [13]. In **Tilde MODEL** corpus there are over 10M segments of multilingual open data for publication on the META-SHARE repository, maintained by the Multilingual Europe Technology Alliance, and on the EU Open Data Portal. The data has been collected from sites allowing free use and reuse of its content, as well as from Public Sector web sites. **WikiMatrix**

CORPUS	Tokens	Links	Sentence pairs (MOSES format)	Words (MOSES format)
Europarl.en-sl	31.5 M	0.6 M	624,803	27.56 M
CCAligned.en-sl	131.3M	4.4 M	4,366,555	110.08 M
DGT.en-sl	215.8M	5.2 M	5,125,455	162.58 M
MultiCCAligned.en-sl	5.6 G	4.4 M	4,366,542	110.01 M
OpenSubtitles.en-sl	178.0 M	2.0 M	19,641,477	213.00 M
TildeMODEL.en-sl	2305.4 M	21.1 M	2,048,216	79.90 M
WikiMatrix.en-sl	1.1 G	0.9 M	318,028	11.99 M
wikimedia.en-sl	350.6 M	31.8 K	31,756	1.50 M
XLEnt.en-sl	200.7 M	0.9 M	861,509	4.53 M

Table 1. Size of datasets for the general translation model.

v1 is a parallel corpus from Wikimedia compiled by Facebook Research. The approach is based on multilingual sentence embeddings to automatically extract parallel sentences from the content of Wikipedia articles in 85 languages. To get an indication on the quality of the extracted bitexts, neural MT baseline systems are trained on the mined data only for 1886 languages pairs, and evaluated on the TED corpus, achieving strong BLEU scores for many language pairs [14]. The **Wikimedia v20210402** corpus contains Wikipedia translations published by the Wikimedia Foundation and their article translation system. The **XLEnt v1** corpus was created by mining CCAligned, CCMatrix, and WikiMatrix parallel sentences. These three sources were themselves extracted from web data from Commoncrawl Snapshots and Wikipedia snapshots. Entity pairs were obtained by performing named entity recognition and typing on English sentences and projecting labels to non-English aligned sentence pairs. The lexical/semantic/phonetic alignment approach yielded more than 160 million aligned entity pairs in 120 languages paired with English.

2.2 Domain translation model

Our domain translation model will be specialized in translating TED Talks. The main goal of this project is to provide a useful and effective tool for translating speech, especially TED Talks, and this way ensuring access to a wide range of talks and other speeches in the Slovene language. We expect some difficulties in training the domain translation model, namely a confusion regarding the word order and terminology because TED Talks address a plethora of different topics and their speech characteristics differ greatly from written texts. Despite being aware of the countless enhancements necessary to provide an excellent end product, our focus in this fairly short time frame is to provide a machine translator that will work effectively and offer comprehensible and useful translations.

For the domain-specific machine training, we opted for the two TED Talk corpora accessible on the Opus website – the TED2013 and TED2020 corpus. The included texts are mainly transcripts of speeches on various topics and their Slovene translations. Both datasets add up to 1.8 million words (MOSES format) and 2.1 million tokens which is enough to form a well-rounded base for machine learning. For more information about the domain-specific corpora see Table 2.

We expanded the datasets by manually aligning 15 TED Talks from 2019 and 2018 that are available on the TED website (https://www.ted.com/talks).

CORPUS	Tokens	Links	Sentence pairs (MOSES format)	Words (MOSES format)
TED2013	0.5 M	15.2 k	14,960	0.45 M
TED2020	1.6M	43.9 k	44,340	1.35 M
Extras	23005	/	983	/

Table 2. Size of datasets for the domain translation model.

3. Methods

3.1 Pretrained model

As a baseline for evaluating our models, we found an already trained model, available in HuggingFace (https://huggingface.co/Helsinki-NLP/opus-mt-en-zls) [4]. It is a transformer-based multilingual model that includes all the South Slavic languages. The framework provides both the South-Slavic to English model and the English to South-Slavic model. On the Tatoeba test dataset for Slovenian, the English to South-Slavic (en-zls) model achieves an 18.0 BLEU score and 0.350 chr-F score.

The model in question was trained using MarianNMT [6]. The authors applied a common setup with 6 self-attentive layers in both, the encoder and decoder network using 8 attention heads in each layer. SentencePiece [7] was used for the segmentation into subword units.

The translation model can be loaded through the *transformers* library in Python and for translation into Slovene, we must add the Slovene language label at the beginning of each sentence (>>slv<<).

3.2 Training from scratch

For training our model from scratch, we have decided to use fairseqs (already described in Section 1.1) extension [15] that

has additional data augmentation methods. We have trained our general model on a corpus described in Subsection 2.1.

3.2.1 Preprocessing

Before training the model, we had to preprocess the data. The datasets were already formatted as raw text with one sentence per line and with lines aligned in English and Slovenian datasets. We first normalized the punctuations, removed non-printing characters, and tokenized both corpora with Moses tokenizer [16]. We removed all the sentences that were too short (2 tokens or less) or too long (250 tokens or more) and the ones where the ratio of lengths was larger than 5 because there is a good chance that sentences like this are not translated properly. We then applied Byte pair encoding (BPE) [17] to the dataset. The algorithm learns the most frequent subwords to compress the data and thus induces some tokens that can help recognize less frequent and unknown words.

With this preprocessed data, we then built the vocabularies that we used for training and binarized the training data. Cleaned and preprocessed training data has $\approx 16M$ sentences with $\approx 345M$ tokens in English and $\approx 341M$ in Slovene. Both of the vocabularies have around 45000 types.

3.2.2 Training

We trained a transformer [18] model. We used Adam optimizer, inverse square root learning rate scheduler with an initial learning rate of $7e^4$ and dropout. We also used the proposed augmentation with cut-off augmentation schema that randomly masks words and this way produces more training data and a more robust translator.

3.3 Fine-tuning on TED talks

The general translation model that we trained from scratch, will be later fine-tuned on domain training data. **TBA**

3.4 Evaluation

In order to test the performance of the general translation model and the fine-tuned translation model for TED Talks, we have to evaluate the results.

The automatic evaluation will be carried out on three validation sets. First, the general translation model will be evaluated on a subset of the general training data (hereinafter referred to as the general validation set). All three models will be evaluated on a subset of the domain training data (hereinafter referred to as the domain validation set), as well as the validation set provided by the assistant (hereinafter referred to as the assistant validation set). Manual evaluation will only be performed on a subset of the domain validation set, as described in 3.4.2.

3.4.1 Automatic evaluation

Since the manual evaluation of the translations is very time-consuming, it is very difficult to evaluate a sufficient amount of sentences this way. In this case the automatic evaluation metrics are often used. Because natural language is quite subjective, the perfect measure does not exist, but by evaluating our results with different techniques, we will be able to assess how successful our translation model is and to compare it with others. In the following subsections, we describe some of the more important metrics that will be used. All of the metrics will be calculated using the *nltk* [19] library in Python.

Bilingual evaluation understudy (**BLEU**) [20] score is the most commonly used score for comparing different approaches for automatic translation. It tests the quality of machine-translated texts by comparing them to a professional human translation. The algorithm operates by counting matching n-grams between the candidate and reference translations where matches are positionally independent. The higher the number of matches, the higher the BLEU score.

Metric for Evaluation of Translation with Explicit ORdering (**METEOR**) [21] was designed to fix some problems found in BLEU. It is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. Compared to BLEU, it achieved a higher correlation with human judgment.

NIST metric is based on BLEU, but with some differences. For example, when calculating n-gram precisions, NIST also calculates how informative a particular n-gram is, meaning that rarer correctly matched n-grams are weighted more, whereas in BLUE, all are weighted equally.

Character n-gram F-score (**chrF**) [22] is based on character n-gram precision and recall enhanced with word n-grams. The tool calculates the F-score averaged on all character and word n-grams, where the default character n-gram order is 6 and word n-gram order is 2. The arithmetic mean is used for n-gram averaging.

GLEU [23] is an SVM-based machine learner, based on a metric that estimates fluency by examining the use of parser outputs as metrics. The authors show that the machine learner provides a consistent estimator of fluency.

Rank-based Intuitive Bilingual Evaluation Score (**RIBES**) is a metric that was developed for distant language pairs. It focuses on the word order which can be different in distant languages. It is based on rank correlation coefficients (specifically Kendall's and Spearman's), modified with precision.

3.4.2 Manual evaluation

The texts will also be evaluated manually, namely by the fluency-adequacy criterion. For this part of the evaluation, the Excel format will be used. Each of the evaluators will be designated particular texts from the domain validation set and will randomly choose 30 segments translated from English into Slovene. If there will be enough time, additional segments will be evaluated. To determine the adequacy of the translation, the evaluator marks how much of the meaning expressed in the source text is also expressed in the target translation. To determine the fluency of the translation, the evaluator marks whether or not the translation is grammatically well-formed, whether it contains correctly spelled words, whether it is intuitively acceptable, and can be sensibly interpreted by a native speaker. To test the adequacy, the evaluator must compare the source text and the translation and, to check the fluency, they can focus merely on the translation. The evaluators give the

answers to both questions on a scale from 1 to 4. We chose this evaluation technique because it describes the quality of the translation clearly and simply. The evaluators will be the translators in our group and other translation students from the Faculty of Arts in Ljubljana.

Because we plan to evaluate three different translation models, we will need to evaluate the same texts three times. Evaluating one text multiple times by the same person is not recommended, therefore, we will switch the translations between the members when starting the evaluation of each translation model.

Finally, we plan to evaluate the domain machine-translated texts from the users' point of view. Evaluators, who are not familiar with this project, will be given the translated text and a questionnaire formed by the translation team in this project. The questionnaire will verify whether the users understand the information given in the text, meaning it will test the translation's functionality in the eyes of the user. We chose this evaluation technique because it shows whether or not the translation is in fact functional and useful. The evaluation results will be interpreted by the translators from our team.

4. Results

In Table 3, we present the results for the pretrained model. Currently, we have been able to train our model for 1 epoch, but we haven't evaluated the results yet.

5. Discussion

Acknowledgments

References

[1] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, "A statistical approach to machine translation," *Comput. Linguist.*, vol. 16, no. 2, p. 79–85, Jun. 1990.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[3] Z. Lin, X. Pan, M. Wang, X. Qiu, J. Feng, H. Zhou, and L. Li, "Pre-training multilingual neural machine translation by leveraging alignment information," 2021.

[4] J. Tiedemann, "The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT," in *Proceedings of the Fifth Conference on Machine Translation*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1174–1182. [Online]. Available: https://www.aclweb.org/anthology/2020.wmt-1.139

[5] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[6] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast neural machine translation in C++," in *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia, 2018. [Online]. Available: https://arxiv.org/abs/1804.00344

[7] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *CoRR*, vol. abs/1808.06226, 2018. [Online]. Available: http://arxiv.org/abs/1808.06226

[8] A. V. Miceli Barone, J. Helcl, R. Sennrich, B. Haddow, and A. Birch, "Deep architectures for neural machine translation," in *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 99–107. [Online]. Available: https://www.aclweb.org/anthology/W17-4710

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 06 2017.

[10] M. Junczys-Dowmunt and R. Grundkiewicz, "An exploration of neural sequence-to-sequence architectures for automatic post-editing," 06 2017.

[11] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *CoRR*, vol. abs/1910.10683, 2019. [Online]. Available: http://arxiv.org/abs/1910.10683

[12] J. Tiedemann, "Parallel data, tools and interfaces in opus," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.

[13] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 923–929. [Online]. Available: https://www.aclweb.org/anthology/L16-1147

[14] H. Schwenk, V. Chaudhary, S. Sun, H. Gong, and F. Guzmán, "WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia," 2019.

[15] stevezheng23, "fairseq_extension." [Online]. Available: https://github.com/stevezheng23/fairseq_extension

Model	Dataset	BLEU	chr-F	GLEU	METEOR	NIST
Pretrained	Assistant	0.33	0.59	0.35	0.51	7.80
	Domain	0.17	0.50	0.21	0.37	5.14
General	Assistant					
	General					
	Domain					
Domain	Assistant					
	Domain					

Table 3. Evaluation scores for all models and all validation datasets.

[16] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 177–180. [Online]. Available: https://www.aclweb.org/anthology/P07-2045

[17] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: https://www.aclweb.org/anthology/P16-1162

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[19] E. L. Steven Bird and E. Klein, *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[20] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: a method for automatic evaluation of machine translation," 10 2002.

[21] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72. [Online]. Available: https://www.aclweb.org/anthology/W05-0909

[22] M. Popović, "chrF: character n-gram F-score for automatic MT evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 392–395. [Online]. Available: https://www.aclweb.org/anthology/W15-3049

[23] A. Mutton, M. Dras, S. Wan, and R. Dale, "GLEU: Automatic evaluation of sentence-level fluency," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 344–351. [Online]. Available: https://www.aclweb.org/anthology/P07-1044