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

Neural machine translation

Marko Cigan, Janja Koželj, Januš Likozar, Miha Šemen, Kristian Wicher, Iva Županc

Abstract

In this report we describe how we trained a neural machine translation model for translating literary text from English to Slovene. We used preexisting frameworks and libraries to try various models, for final evaluation we chose a transformers based model. We first trained our model on general data and then finetuned it on domain specific data. We also tried to finetune model on texts from specific authors to see if it makes an improvement. We evaluated translated texts automatically and manually. The results show that finetuning improves translations of literary texts.

Keywords

neural machine translation, Slovenian-English, domain specific translation, literary texts

Advisors: Slavko Žitnik, Špela Vintar

Introduction

Machine translation is a field that connects computer science with linguistics. The most basic approach is replacing words one by one with its translation, but these translations are rarely good. More complex methods such as statistical machine translation (SMT) use probabilistic models to generate translations based on the analysis of corpora from both languages.

In this project we use neural machine translation (NMT) methods that use artificial neural network to generate translations. Compared to more basic approaches to translation it does not need any knowledge about the languages, like grammatical structure or similar. NMT methods use one neural network that is trained end to end on source and target language corpora. The corpora used has to be aligned, so that sentences match between both languages.

Earlier NMT methods use encode-decoder recurrent neural networks (RNN), which account for sequential data, so this style of networks can also be used for sequence to sequence prediction tasks in other fields. These methods use multiple Long Short-Term Memory (LSTM) blocks to encode input vectors and decode target output vector, one such method is described in [1]. It maps input sequences to fixed internal representation, but still manages to translate longer sentences as well. To address problems with rare words and lack of robustness, in [2] they add attention layers to similar encode-decoder architecture as described before.

Latest state of the art results use transformer based neural networks, which rely on self attention mechanisms to compute internal representation of inputs and outputs [3].

Related work

There is lots of research in NMT area to create general translators, but not so many that target translating literary texts, especially novels. In [4] Toral and Way compare classical statistical MT methods with NMT, both adapted to literary domain, on English to Catalan translations. They show performance improvements of NMT over SMT methods by comparing BLEU scores, but also additionally human evaluation, where 17-34% of translations made by NMT were considered of equivalent quality to professionally made translations.

To describe quality of NMT models translations in [5] Fonteyne et al. do a document level evaluation of a novel translated from English to Dutch using GNMT [2]. In the translated text they look for all types of fluency and adequacy errors and report that 44% of the sentences do not contain any errors, with the most common type of mistake being mistranslation.

The task of training and evaluating a NMT model from English to Slovene is described in [6], where Kuzman et al. compared GNMT and bespoke NMT models, later one adapted to perform on literary texts. They report that general GNMT outperforms all of the other specific models, but also specific models tailored to certain author outperform models trained on a more diverse literary dataset.

Methods

We made our code available at https://github.com/JanusL/nlp-project.

NMT Frameworks

There are many NMT frameworks that make training and inference of translation models easier. Below we list some of the most widely known ones:

- **MarianNMT** [7] is NMT toolkit written in C++. It is made to be fast and self-contained. It offers training and translating using some common models, but has a bit lacking documentation.
- **OpenNMT** [8] is a NMT toolkit that has TensorFlow and PyTorch version. It supports translation, summarization and other tasks, and includes tools for fast inference and tokenization. It is made to be easy to use with many examples and lots of documentation.
- **Trax** [9] is a library for deep learning using TensorFlow that supports training of varous common types of networks and reinforcement learning algorithms for translation and other natural language processing tasks. It is a successor to seq2seq toolkit [10], but does not have the same extensive support and documentation yet.
- **Fairseq** [11] is a sequence modeling toolkit written using PyTorch that supports translation, summarization, language modeling and other text generation tasks. It offers lots of pretrained models, models and examples for lots of specific tasks and has extensive documentation and support.
- **Transformers** [12] is a natural language processing library that uses PyTorch, TensorFlow and Jax. It contains a lot of pretrained models, model configurations and other tools for various tasks. It offers extensive documentation and support with many examples, but it is not command line end to end tool.

Considering our use case we chose to use OpenNMT toolkit, the PyTorch version, because it is easy to use, has many examples and good documentation. We also used Transformers library for tokenization.

OpenNMT toolkit can be used as a python library or from the command line using scripts and configuration files. There are three main features, building vocabulary, training the model and model inference (translation of text). For scripts there has to be a configuration file (or many command line arguments), which defines all the parameters. For use in OpenNMT we have source (in our case English) and target (Slovenian) text files of aligned sentences, with one sentence per line.

When building vocabulary we can set maximum number of tokens in vocabulary. We can have separate vocabulary for each language or one that is shared by both. By default it takes input as words separated by spaces, but we can set it to use a tokenizer. There are two tokenizers available, byte-pair encoding (BPE) and SentencePiece.

For training we can define custom models, there are many available architectures for encoder and decoder parts, like RNN, CNN (convolutional neural networks), Transformers and other variations, we can also define attention and self attention blocks. We can also set many parameters for each building block in architecture. It supports use of pretrained embedding and training on multiple gpus.

For translation we can choose any trained model to generate translations of a given text.

Data

We decided to make a NMT translator for literary texts. There are multiple approaches how to train a model that is domain specific [13]. One is to train on only domain specific data, which is difficult if there is little appropriate available corpora, so the quality of model depends on chosen domain. Another approach is to train on all available data that is not necessarily domain specific and then fine tune the model on specific corpora. Since there is not much already prepared available corpora of literary texts to train model from scratch, we used more general corpora for initial training.

Translation of literary texts is a difficult task since the source text is lexically rich and literal translation is not necessarily the best. [5, 4, 14]. In case of translation from English to Slovenian there is also a challenge of Slovenian being morphologically complex language and the small number of available aligned corpora of these two languages [6].

For training the general NMT model we used the provided TC3 dataset. Later we realized that the amount of data was insufficient. So, we added other resources either through databases provided in class or on the internet. We focused on non-specific and non-technical texts, but also used the open parallel corpus Opus as a source to find bilingual corpora. As a result, we used the following files and corpora to teach our NMT model:

- TC3 - includes EUparl, OpenSubtitles 2018, EMEA - European Medicines Agency, DGT - translation memories and ELRC - statistical reports. (24419755 sentence pairs)
- TRANS3 – legacy parallel EN-SL corpus of texts from different domains (engineering, health, legal); now integrated into Evrokorpus. [15] (16160 sentence pairs)
- Various TMX from the IT field (provided by prof Vintar): various aligned texts from different sources in the field of informational technology (created within a project for the subject Korpusi in lokalizacija (Corpora and localization) at MA Translation at the Faculty of Arts of the University of Ljubljana). It includes

abstracts of MA thesis, abstracts from articles of the science magazine Uporabna informatika, various instruction for appliances, instructions and tips for software, some webpages and other texts. (22956 sentence pairs)

- Wikimedia - general texts about various topics. [16] (31755 sentence pairs)
- Wikipedia - general texts about various topics. [16] (140124 sentence pairs)

We believed that teaching our NMT model on literary texts would be much easier than on other technical texts and provide us with more accurate translation results, which we could then later compare, improve and evaluate. For training and evaluating our model on domain specific data we used the following corpora:

- SPOOK - a dataset that was created within the translational project Slovensko prevodoslovje: viri in raziskave (under the project code J6-2009). The corpus that was created, called the Slovenian translational corpus (SPOOK), is a multilingual parallel corpus comprised of Slovenian translations and the original works in English, German, French and Italian. The creation of this specific corpus was financed by the Slovenian Research Agency. We used material from the two subcorpuses of SPOOK which contain the English originals and their Slovenian translations. The two subcorpuses are made of 9 literary texts each, together 18. These 9 novels and their translations are *The Supernaturalist, The Way Through the Woods, Harry Potter and the Half-Blood Prince, Harry Potter and the Deadly Hallows, The Lord of the Rings: The Two Towers, The Curious Incident of the Dog in the Night-Time, White Teeth, The Da Vinci Code, The Fifth Child*. The novels are from the fantasy (three novels), mystery (three novels), science fiction (one novel), horror (one novel) and realism (one novel) genre. [17] (67920 sentence pairs)
- George Orwell's Nineteen Eighty-Four - a translated book from ELAN dataset, that we used for evaluation. [18] (6639 sentence pairs)

Data preprocessing

First we converted all data to a format that is expected from OpenNMT toolkit training script, two text files, one for each language, with one sentence per line. We checked all matched sentences so that there were no missing ones in any language, and checked that they matched. Then we cleaned the text so that we had uniform symbols across different corpora, for example we switched occurrences of *&quot* with *"* and similar.

Tokenization and Vocabulary

We have tried different tokenization methods, like WordPiece and SentencePiece [2]. Both are subword tokenization algorithms, WordPiece generates tokens from characters in each word separately and SentencePiece generates tokens from subword units from whole sentence. SentencePiece can be trained end to end on raw sentences, but in our case we used WordPiece since we already had available pretrained version on much more data, than what we had available.

WordPiece is a subword algorithm where we start with individual characters and merge them together based on most frequent combination of characters in text. To calculate WordPiece tokens we used BERT tokenizer from Transformers library with pretrained vocabulary from CroSloEngual BERT model [19]. It contains 49601 tokens.

Embeddings and models

Inputs to the models are vectors that we get by embedding input tokens. In OpenNMT toolkit the embeddings are learnt during training of the model. We tried different types of models, like RNN and Transformers based ones.

Evaluation

For evaluation we will use automatic evaluation methodologies like BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit ORdering), CHRF (Character n-gram F-score), GLEU (Google BLEU) and NIST (n-gram co-occurrence statistics). They determine quality of machine translated text compared to the true translation. The implementations for score calculation are available from nltk (natural language toolkit) library.

Besides automatic evaluation we also manually evaluated the translated texts. We decided to use the adequacy and fluency method to evaluate the translations during different stages of the training [20]. This is one method used to judge an MT's output. It is divided into two parts, fluency measures how much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation, regardless of the correct meaning, while adequacy measures whether the translation conveys the correct meaning, even if the translation is not fully fluent. We chose this method due to the NMT being more suited to translating short syntactic structures i.e. sentences rather than whole paragraphs [21].

The translation results are then evaluated using grades ranging from 1 to 4, where in the fluency category 1 means that the translated text is incomprehensible for a native speaker and 4 being a flawless simulation of a native speaker. Adequacy is measured similarly from 1 to 4. 1 denotes almost none of the originals meaning was conveyed and 4 is used for translations that are fluent and include everything from the source text.

Experiments

Simple LSTM model

For training, we use the command line approach with a configuration file. First we convert the SPOOK dataset from the .xml files into a source and target text files, containing aligned sentences in English and Slovenian. We do the same for the Orwell dataset, which we use for evaluation.

We create a training.yaml file, which we will use to store all our parameters. First we define our data and their paths, and the path where we want to save our vocabulary. We also define SentencePiece tokenization and the path where the SentencePiece model will be saved. When these parameters are defined, we can run `onmt_build_vocab -config training.yaml` to build our vocabulary and tokenization model. This resulted in two separate vocabularies for each language, each containing 16000 tokens.

Before we start training, we define an Adam optimizer with a learning rate of 0.001, and set our batch type to tokens (default is sentences) and batch size to 1024. Since we intend to use the default LSTM network with 2 layers and 500 units on both encoder and decoder, we do not have to define it. We also set number of steps to 100000. We then run `onmt_train -config training.yaml` to start training. We can also use TensorBoard to make it easier to view how the model is training.

To look at how well our model is translating, we can use the `onmt_translate` script, but it does not support tokenization, meaning we need a separate script to tokenize our data with the SentencePiece model, use the script to translate it, then use another script to detokenize it before we can evaluate it. It is easier to setup an translation server with `onmt_server`, where we can then send translation requests and receive the translations using http requests with curl.

Transformer models

To increase performance of our model, we use the state of the art Transformer architecture, which uses an attention mechanism to process sequential data. We use a model with 6 encoder layers, 6 decoder layers and 8 attention heads and dropout of 0.3. For word embeddings, we use 512-dimensional vectors, and we share the embeddings between our decoder and encoder. For the optimizer, we use the adam optimizer with noam weight decay, learning rate of 2 and 8000 warmup steps.

For use with transformers, the SPOOK dataset is far too small and quickly causes overfitting. We first train a general NMT model on the provided TC3 dataset and evaluate it on provided test dataset.

Instead of training our own SentencePiece model, we use a pretrained BERT WordPiece model, which was trained on far more data and should be more representative of the language. Since OpenNMT does not support WordPiece tokenization, we use a script to tokenize our data prior to building our vocabulary. We also clean up the data by replacing all symbols with the same meaning with one symbol, such as using " for all types of quotation marks, in order to reduce the number of unknown symbols in our data. We also provide a script to shuffle the training sentences. Path to our cleaned, tokenized and shuffled data is then written to the configuration file.

We then train the model for 100 000 steps, which results in the translations provided in the intermediate results table.

Trained models

After initial test trainings, four main models were trained and evaluated, including a general model and three models finetuned on literary texts. Some additional transforms were added to the data preprocessing as well: all sentence pairs where one pair is longer than 150 tokens or one is longer than the other by more than 50% were removed.

- General transformer (**GT**): For the general model, we use the same architecture as described above. We add Wikimedia, Wikipedia, Random IT texts and Trans3 to the training data. The model was trained for 300 000 steps.

- Finetuned on Spook (**GTS**): As SPOOK does not contain enough data to train a model on its own, the general model was finetuned. After the 300 000 steps training on general data, we train it for 20 000 additional steps on SPOOK data only.

- Finetuned on Harry Potter (**GTS-HP**): The general model was trained of 20 000 aditional steps on the Harry Potter novel from SPOOK only.

- Finetuned on Lord of the Rings (**GTS-L**): Instead of finetuning on LOTR from SPOOK only, we add other data to training in order to prevent overfitting. The data consisted 1/3 of LOTR, 1/3 other SPOOK texts and 1/3 general texts.

Evaluation process

To limit the variance that could be created through 4 different evaluators, we first selected 10 random sentences, which each evaluator then rated on their own. We did this to check how each of us evaluates the translation, what criteria each of us set and how each of us understands a certain score (flawless, good, disfluent, incomprehensible or everything, most, little, none). After we had all evaluated the ten sentences, we realized that we did not have a common understanding of the fluency instructions exactly and that our evaluations varied heavily. We discussed our results, shared our understanding of the instructions and matched the way we evaluated the translations as much as possible.

After reevaluating the 10 sentences we got a maximum fluency rating of 2.8 and an adequacy rating of 2.3. The minimum rating for both categories respectively were 1.6 and 1.5. Therefore, we could expect a minimal +/– 0.6 differential in fluency and a +/– 0.4 in adequacy rating from the median. The average rating was 2.3 for fluency and 1.8 for adequacy.

Later we evaluated four files with the translations of the same sentences from Orwell's novel 1984, but from four different iterations of the NMT. We decided that we would all evaluate the first 50 translated sentences from the files. The evaluation was done independently, but we pointed out various errors from the translations to see how others would rate them.

Our final evaluations were a test to see how well the NMT would be able to translate abstracts from a shared literary universe as a novel from SPOOK. For this we decided to let it translate different abstracts from fanfiction of Harry Potter found on the internet [22, 23, 24, 25]. The abstracts were translated using three different iterations of the NMT (the general model, a finetuned version on the database SPOOK and a finetuned version on the two Harry Potter novels in SPOOK). The total count of these translated sentences was 30. Even though, only 3 of the 4 evaluators evaluated these translations, the minimal differentials remain the same, due to the minimum and maximum evaluators being represented within this group.

Results

Train time

To train the models, we used a single Nvidia RTX 2070 GPU with 8 GB of memory. Training took 1 hour per 10 000 steps and training our general model took 30 hours over 3 days. Finetuning took aditional 2 hours per model.

Parameters

During testing, many diffrent parameters and preprocessing steps were tried with varying success. As training a model takes a long time, we were severely limited in how many parameters we could try.

- Removing sentence pairs where one sentence is 50% longer than the other: This step helps in removing sentence pairs where one sentence contains more content than the other, which can reduce the performance of the model. In testing, using this transform has proven to help prevent translations, where only part of the sentence was translated and the other part would be ignored.
- Lower casing data: A model was trained with all words lowercased before tokenization to try to increase robustnes and to prevent upprecased words taken diffrently than same words, but lowercased. This ended up performing worse, as the model had a hard time understanding names.
- Using seperate embeddings: Instead of using shared embeddings for both encoder and decoder, using seperate embeddings was tested. This had promising results with some words translating better, but we were out of time for further testing.
- Using some general data while finetuning: Since we are finetuning on a very small ammount of data, the model can quickly overfit and ”forget” structure that it has learned on general data. To try to improve this, we added some SPOOK and some general data to training of our LOTR model. This has seemed to have some improvements, but exact weights of each corpus would have to be determined for best result.

Automatic evaluation results

The results of automatic evaluation of the models on given general test set can be seen in Table 1 and on Orwell's Nineteen Eighty-Four in Table 2. The general model achieves quite good scores on general test set, but much worse scores on literary texts. The model GTS finetuned on SPOOK data achieves slightly better scores on Orwell's Nineteen Eighty-Four than general model. The low scores, especially on literary texts can be a consequence of model not translating words exactly the same as they are in original translation, but using synonyms or words with similar but not exactly the same meaning.

The finetuning models on certain authors book also decreases model performance on literary test set, which is probably a consequence of fitting to a more specific style, as LOTR and HP both belong to fantasy genre. Unfortunately we could not perform automatic evaluation on LOTR and HP data to evaluate their finetuned models due to not having aligned official translations of taken abstracts which we used to generate example translations.

There are examples of translated texts in Tables 5, 6, 7, 8. We can see that literary translations of general model often miss names and meanings of words. Especially difficult in that way is Silmarillion text, but after some finetuning it the translations improve. It still does not convey the meaning of sentences very well as we could expect from these kind of long and complicated sentences. In general the model has problems with longer sentences, but that may also be a consequence of training data we used.

Another shortcoming of our method is also not recognising important names as they are very common in literary works, compared to some other domains. This is most apparent in LOTR and HP where if finetuned the model learns what the translations for the names are, determined by professional translators, otherwise it just uses the English name and declines it to fit into sentence.

model	BLEU	CHRF	GLEU	METEOR	NIST
GT	**0.1165**	**0.3184**	**0.1593**	**0.3184**	**4.6743**
GTS	0.0043	0.0778	0.0287	0.0778	1.0823
GTS-L	0.0118	0.0921	0.0317	0.0921	1.1849
GTS-HP	0.0000	0.0260	0.0083	0.0260	0.0022

Table 1. Results of NMT models evaluated on provided test set.

model	BLEU	CHRF	GLEU	METEOR	NIST
GT	**0.0024**	0.2611	0.0339	0.2611	0.0088
GTS	0.0013	**0.2887**	**0.0358**	**0.2887**	**0.1069**
GTS-L	0.0010	0.2815	0.0336	0.2841	0.0207
GTS-HP	0.0003	0.2369	0.0287	0.2369	0.0155

Table 2. Results of NMT models evaluated on Orwell's Nineteen Eighty-Four.

Manual evaluation results

From the results in Table 3 it is clear to see based on the larger evaluation sample from the NMT translations of the novel 1984, that the model improved in adequacy and fluency with almost every iteration. Factoring in the minimal differentials between the different evaluators, would also explain the lower final rating, that still shows an improvement from the general model.

The results of manual evaluation of abstracts from Harry Potter fanfiction are in Table 4. In the smaller sample size from the 30 translation examples from Harry Potter fanfiction the biggest difference in ratings seems to be with the model finetuned on the SPOOK corpora. This could again be explained through evaluation discrepancies since the same evaluator also evaluated the final version of the 1984 NMT and also gave lower ratings for fluency and adequacy.

Some general observations were that the NMT performed much better when translating shorter sentences, often scoring a 4 in fluency, but would have much more issues in adequacy. Overall, an improvement was seen with every iteration in both analyzed sample sizes. The main difference being the elimination of common mistakes with capitol letter use, repeating the same word over and over, not knowing how to translate symbols such as quotation marks, etc. The fanfiction translation also had better results due to the source text containing less complex sentences. The example also showed the NMT being capable of translating single sentences of works in the same literary universe as the novels in the database corpora.

Model	Adequacy	Fluency
GT	1.36	2.28
GTS-5000	1.84	**2.76**
GTS-10000	**1.96**	2.66
GTS-15000	1.54	2.40

Table 3. Results of manual evaluation of models on Orwell's Nineteen Eighty-Four. GTS models have number of finetuning steps written next to them.

Model	Adequacy	Fluency
GT	**2.13**	2.87
GTS	1.53	2.17
GTS-HP	2.07	**3.37**

Table 4. Results of manual evaluation of models on Harry Potter fanfiction examples.

Discussion

We observed improvements in the future iterations of the NMT and in the NMT models, which were finetuned to better fit the literary cannon from the source text. To prove whether a future model would be capable of translating other novels or texts from the same literary universe into more adequate and fluent solutions would require more time. The limited timeframe given for the model to learn also limits our capabilities of judging it's ceiling.

We observed that certain mistakes were gone with future iterations, but the question whether a NMT based on a total of nine novels and their translations, can produce much more adequate solutions remains unanswered. To get better results we would need a much larger literary corpora, preferably of such size that it could replace the general training data we used, so that the model would be trained on only domain specific data. Using more (literary) data would also allow us to train the model for many more iterations, compared to now, which would probably give us better translations. We tried to finetune model on works that were from same author (Tolkien) or very similar in content (Harry Potter) and we got good results, but we can not say that it helped more that finetuning on SPOOK, due to the finetuning data already being present in the SPOOK dataset.

Conclusion

We trained a NMT translator adapted to literary texts. The main shortcoming and a place for improvement is the amount of training data, especially domain specific parts. By using more literary data for training, the fluency and adequacy of translations should improve. We also did not have enough time or computing power to search over the model parameters to find optimal settings.

For future research we also propose a more genre-based approach to text-selection, since in our case we had various genres present in the SPOOK dataset. We would also need to experiment with more diffrent parameters in order to find the best model. For finetuning, we would also have to find a good way to reduce overfitting, such as finding a good amount of general data to include in training.

References

[1] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.

[2] 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.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[4] Antonio Toral and Andy Way. What level of quality can neural machine translation attain on literary text? In *Translation Quality Assessment*, pages 263–287. Springer, 2018.

English	In the absence of measures, a further deterioration in the Union industry's situation appears unavoidable.
Slovene	Brez ukrepov se zdi, da se nadaljnjemu poslabšanju položaja industrije Unije ne bo mogoče izogniti.
GT	V nenazadnje se zdi, da se položaj industrije Unije slabša.
GTS	V odsotnosti nadaljnjih ukrepov Unije se zdi, da se položaj industrije poslabša.
English	if Article 7(1) applies, the Sanctions Committee has been notified by the Member State of the lien or judgment.
Slovene	če se uporablja člen 7(1), je država članica o zasegu ali sodbi obvestila Odbor za sankcije.
GT	člen 7 (1) se uporablja, če je država članica uradno obvestila Odbor za sankcije ali Odbor za sankcije.
GTS	če je država članica obvestila Odbor za sankcije ali Odbor za sankcije, se uporablja člen 7 (1).
English	Products originating in one of the countries referred to in paragraphs 1 and 2 which do not undergo any working or processing in Turkey, shall retain their origin if exported into one of these countries.
Slovene	Izdelki s poreklom iz ene od držav, omenjenih v odstavkih 1 in 2, ki niso obdelani ali predelani v Turčiji, ohranijo svoje poreklo, če se izvozijo v eno od teh držav.
GT	če se izdelki iz odstavka 1 in 2 ne izvažajo v eno od držav s poreklom iz Turčije ali Turčije, ohranijo svoje poreklo v eno od teh držav.
GTS	če se izdelki iz odstavka 1 in 2 ne izvažajo v eno od teh držav s poreklom iz Turčije, ohranijo ali izvažajo svoje poreklo v eno od teh držav.

Table 5. Comparison of model translations of general test set.

English	The hallway smelt of boiled cabbage and old rag mats .
Slovene	Veža je smrdela po kuhanem zelju in starih , cunjastih predpražnikih .
GT	Stara blazina smrdi po zelju in zelju.
GTS-15000	Stari žimnici so zaudarjali in izsesali smrdljivo dvorano.
English	" Big Brother is watching you " , the caption beneath it ran .
Slovene	" Veliki Brat te opazuje " , se je glasil napis pod njim .
GT	" ; brat Bigath je pod ujetništvom. " ;
GTS-15000	" Veliki brat, " je tekel pod jetnico, ki si jo opazoval.
English	It was curious that he seemed not merely to have lost the power of expressing himself , but even to have forgotten what it was that he had originally intended to say .
Slovene	Čudno je bilo , da mu je ne samo zmanjkalo moči , da bi se izrazil , temveč je celo pozabil , kar je sprva sploh nameraval povedati .
GT	Sprva se je zdelo čudno, da je pozabil povedati, kaj je mislil povedati, vendar ni imel moči, da se je izgubil.
GTS-15000	Kazalo je, da je to tisto, kar je sprva hotel povedati, vendar ni pozabil, da je izgubil moč, da bi izrazil radovednost.

Table 6. Comparison of model translations of Orwell's Nineteen Eighty-Four.

[5] Margot Fonteyne, Arda Tezcan, and Lieve Macken. Literary machine translation under the magnifying glass: Assessing the quality of an nmt-translated detective novel on document level. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3790–3798, 2020.

[6] Taja Kuzman, Špela Vintar, and Mihael Arcan. Neural machine translation of literary texts from english to slovene. In *Proceedings of the Qualities of Literary Machine Translation*, pages 1–9, 2019.

[7] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia, 2018.

[8] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[9] Trax - Deep Learning with Clear Code and Speed. https://github.com/google/trax. [Online; accessed 20.5.2021].

[10] D. Britz, A. Goldie, T. Luong, and Q. Le. Massive Exploration of Neural Machine Translation Architectures. *ArXiv e-prints*, March 2017.

[11] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[12] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chau-

English	There was Eru, the One, who in Arda is called Ilúvatar; and he made first the Ainur, the Holy Ones, that were the offspring of his thought, and they were with him before aught else was made.
GT	Eden od njegovih potomcev je bil Aiu Ardu, ki so ga sestavljali še preden so mislili, da je bil z Ainu, Svetom, Ervatarjem, Ilarjem in Svetom.
GTS	Eden od njih je bil tisti, ki so ga imenovali Éu Ilvatar, Éu, Éu, Ajnugúr, Éu in Éu, ki je bil v misli, da so bili sestavljeni iz potomcev, ki so jih naredili pred njim, in je bil prvi, ki je bil v svoji auuuuju, je bil v mislih.
GTS-L	= = = Življenje in delo = = = Eden od Aiu Ardúnu, ki je bil pred njim, je bil prvi potomec Sv. Ergúnu, ki so ga sestavljali skupaj z Aiu, Il, ki je bil narejen v misli, da so potomci Sv.
English	And of these Melkor was the chief, even as he was in the beginning the greatest of the Ainur who took part in the Music.
GT	Melkornu je bil vodja glasbene šole Ainu, ki je bil v Andkorju največji del teh časov.
GTS	In to je bil tudi najbolj del zadnjega dela Glasbo v Ainukornuju, ki je bil najpomembnejši izmed vseh tistih, ki so ga v začetku leta jemali v Jazbino.
GTS-L	In na začetku je bil celo največji izmed teh, ki so ga zasedli v Aikoru, Melnu.
English	Thus it came to pass that of the Ainur some abode still with Ilúvatar beyond the confines of the World; but others, and among them many of the greatest and most fair, took the leave of Ilúvatar and descended into it.
GT	Tako je prišlo do številnih potomcev Ainuja in drugih, vendar je trajalo, da je večino zapustil Ilvatar in jih z vilúdepotar še vedno pustil na cedilu.
GTS	Tako je prišlo do tega, da so se mnogi izmed drugih spuščali v Ajzenúnur, nekateri pa so prišli do tega, da so šli prek Ilvatarja ; nekateri pa so še kar najbolj pošteni in pošteni in pošteni in krepki potomci, ki so se spuščali v Ajzenúr, ostali pa so še vedno z največjim prelazom na svetu.
GTS-L	Tako je prišlo med večino drugih v Ajzenúnu in Ilvatar ; vendar pa je še vedno trajalo, da so mnogi izmed njih potrdili potomce Ajzenadarúnesa in največje prelepega sveta.
English	And therefore they are named the Valar, the Powers of the World.
GT	Zato se imenujejo valenčna valižanska valja.
GTS	Zato se imenujejo tudi Valárji, svetovna sila.
GTS-L	Zato so imenovani " Power Valar ", World.

Table 7. Comparison of model translations of J. R. R. Tolkien's The Silmarillion.

mond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[13] Chenhui Chu and Rui Wang. A survey of domain adaptation for neural machine translation. *arXiv preprint arXiv:1806.00258*, 2018.

[14] Evgeny Matusov. The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, 2019.

[15] Evrokorpus. https://evroterm.vlada.si/info/evrokorpus. [Online; accessed 20.5.2021].

[16] 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).

[17] Slovenian translational corpus. http://nl.ijs.si/spook/. [Online; accessed 20.4.2021].

[18] Tomaž Erjavec. Informatica 26 (2002) 299-307@ aa compiling and using the ijs-elan parallel corpus tomaž erjavec department of intelligent systems jožef stefan institute. *Informatica*, 26:299–307, 2002.

[19] M. Ulčar and M. Robnik-Šikonja. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In P Sojka, I Kopeček, K Pala, and A Horák, editors, *Text, Speech, and Dialogue TSD 2020*, volume 12284 of *Lecture Notes in Computer Science*. Springer, 2020.

[20] Evrokorpus. https://www.systransoft.com/systran/translation-technology/neural-machine-translation-nmt/. [Online; accessed 20.5.2021].

[21] Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. Fluency, adequacy, or hter? exploring different human judgments with a tunable mt metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, 2009.

[22] Snape's Not Dead. https://www.fanfiction.net/s/13751215/1/Snape-s-Not-Dead. [Online; accessed 20.5.2021].

English	The Battle of Hogwarts coming to a pause in the distance, Severus Snape lay slumped against a dented wall in the Shrieking Shack.
GT	Bitka proti zobozdravniku v bitki pri Hogwarju se je bližala stena kačjega blata.
GTS	Robaus Raws je ležal v daljavi, v daljavi, ki je prihajala proti zobozdravniku, ki je prihajala v bitki proti Robausu v Meryascoveeni.
GTS-HP	Bitka proti Rawsu je ležala na zidu, pri čemer je z zobozdravnikom Robausom v bitki za Bradavičarko tik pred Rawsom.
English	Voldemort's chilling telepathic message to gather the dead barely penetrated his flickering mind.
GT	Voldemska telepatska telepatska misel je komajda zbežala, da bi zbežala od mrtvih.
GTS	Mrlakensteinu se je komaj zazdelo, da je Mrlakenstein komaj zaznaval telepatski um in komaj zaznaval razvozlavanje penastega sporočila.
GTS-HP	Mrlakenstein je komaj zaznaval, kako je Mrlakenstein zbegano mrtvo šinilo telepatski um, ki je zbežal.
English	She was on the platform with the scarlet steam engine, glistening with golden letters advertising the fact that it was the Hogwarts Express.
GT	Bila je na oglaševalski platformi Hogen z zlatimi črkami, ki se je bala pred nihanjem.
GTS	Bila je na mesečini, z zlatimi pikami, ki so oglaševali pisma, ki so se svetlikale na trgu Express.
GTS-HP	Z dejstvom, da se je na Bradavičarki izdejanila pisma, je bilo dejstvo, da je oglaševanje na zlatih platformah.
English	A young healer was sitting at the reception counter quietly reading a mag about a possible cure for lycanthropy on the horizon in Witch Weekly.
GT	Sprejemnik je bil čim hitrejši pri branjenju lisic na horizontu, kjer je sedel mladi Witchler.
GTS	V recepciji je bilo mogoče slišati, da je mladi čarovnik na hitro sedel na pultu na pultu in bral pismonoše na obzorju.
GTS-HP	Na mizi je sedel mladi čarovnik s tihim zaznavanjem, da je slišal odgovor na morebitne klice v roki, in sicer je prosil za ponaredek.

Table 8. Comparison of model translations of abstracts from Harry Potter fanfiction.

[23] The Last Crouch. https://www.fanfiction.net/s/13721498/1/The-Last-Crouch. [Online; accessed 20.5.2021].

[24] Halle Potter and the seven years of madness. https://www.fanfiction.net/s/13851500/2/Halle-Potter-and-the-seven-years-of-madness. [Online; accessed 20.5.2021].

[25] The Survivors. https://www.fanfiction.net/s/13875744/1/The-Suvivors. [Online; accessed 20.5.2021].