Neural Translation Model Specialized in Translating English TED Talks into Slovene

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

* Faculty of Computer and Information Science
University of Ljubljana
Večna pot 113, 1000 Ljubljana
{eb1690,es6317}@student.uni-lj.si

† Faculty of Arts
University of Ljubljana
Aškerčeva 2, 1000 Ljubljana
{th3112,mj6953,ss6483}@student.uni-lj.si

Abstract

In this paper, we present our work on a neural translation model specialized in translating English TED Talks into Slovene. The aim is to provide transcriptions of the speeches in Slovene to make them available to a wider audience, possibly with the option of automatic subtitling. First, we trained a transformer model on general data, a collection of corpora from the Opus site, and then fine-tuned it on a specific domain which was a corpus of TED Talks. To see the functionality of the model, we carried out an evaluation of the pretrained, general, and domain versions of the model. We evaluated the translations with automatic metrics and manual methods – the adequacy/fluency and the end-user feedback criterion. The analysis of the results showed that our translation model did not produce the expected results and it can not be used to translate speeches in real life. However, in the TED talks addressing more everyday issues and using simple vocabulary, the translations successfully conveyed the main message of the speech. Any further research should consider improvements, such as including more specialized data covering only one specific topic.

1. Introduction

In this paper, we trained a transformer model from scratch on a large general corpus, which we then fine-tuned on a corpus consisting of TED Talks in order to make a model specialized for the translation of transcribed speeches. We also found a pretrained model for the baseline to which we were able to compare our translation models. We then automatically and manually evaluated all three models on the validation datasets constructed from TED Talks. Finally, we evaluated the general translation model on the validation dataset constructed from the large general corpus.

In Section 3, 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 5 and 6, we present the results and discuss them.

1.1. Goal of the paper

The main goal of this project is to provide a useful and effective tool for translating and subtitling speeches from English to Slovene, and this way granting access to a wide range of talks and other speeches to the Slovene-speaking audience. This paper focuses on translating TED Talks, a form of learning and entertainment that has gained popularity in recent years. Since TED Talks are currently subtitled by volunteer translators, enabling automatic subtitles would facilitate this process. Machine translation (MT) 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. This project was our attempt at machine translation of spoken language, which, if efficient, could also be used for automatic subtitling in general.

2. Related work

There are three main approaches to solving the MT problem, all with their own advantages and shortcomings. The rule-based machine translation (RBMT) is the oldest of the bunch and it requires expert knowledge of both the source and the target language in order to develop syntactic, semantic, and morphological rules. 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 (Brown et al., 1990) is, if given a sentence in the target language, we seek the original sentence from which the translator produced it. Today, as with many computer science fields, the current state-of-the-art approaches for machine translation are based on neural networks. The biggest challenge when building a successful English to Slovene (or vice-versa) automatic translator is obtaining a sufficiently large 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 to pre-training a network on monolingual data (that can be obtained easily) have been proposed.

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) uses two strategies to deal with the problem, namely masked language modeling (MLM) and next sentence prediction (NSP). By using these two strategies in our models, we generally achieve bigger datasets and a model with more context-awareness.

In 2020, the mRASP (Lin et al., 2021) was introduced. Its authors built a pretrained NMT model that can be fine-tuned for any language pair. They used 197M sentence pairs, which is considerably more than we could obtain for only English-Slovene translations.

Although these methods have proven to be successful, one of the largest currently available databases of pretrained translation models was trained using just a standard transformer model and it still achieved great results. The Tatoeba Translation Challenge (Tiedemann, 2020) 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, pretrained translation models for multiple languages were also released and are being regularly updated.

3. Dataset

3.1. General translation model

The datasets for the general translation model are the eight biggest corpora from the Opus site (`https://opus.nlpl.eu` (Tiedemann, 2012)) for the Slovene-English language pair. 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, such as **Europarl**, which is a parallel corpus extracted from the proceedings of the European Parliament from 1996–2011, and the **DGT** corpus, which is a collection of translation memories from the European Commission's Directorate-General for Translation. The other corpora are a collection of translations from different Internet sources, which makes them less reliable, however, they are still very valuable because they ensure a large quantity of the data. These include the **CCAligned** corpus consisting of parallel or comparable web-document pairs in 137 languages aligned with English, the **MultiCCAligned v1** multi-parallel corpus, the **OpenSubtitles** corpus compiled from an extensive database of movie and TV subtitles, the **Tilde MODEL** corpus consisting of over 10M segments of multilingual open data for publication on the META-SHARE repository, the **WikiMatrix v1**, a parallel corpus from Wikimedia compiled by Facebook Research, the **Wikimedia v20210402** corpus, and the **XLEnt v1** corpus created by mining CCAligned, CCMatrix, and WikiMatrix parallel sentences. The exact size of each one, complete with the number of tokens, links, sentence pairs, and words, is noted in Table 1.

3.2. Domain translation model

Our domain translation model is specialized in translating TED Talks.

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 2018 and 2019 that are available on the TED website (https://www.ted.com/talks).

4. Methods

4.1. Pretrained model

As a baseline for evaluating our models, we found an already trained model, available in HuggingFace (Tiedemann, 2020). 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 Slovene, the English to South-Slavic (en-zls) model has achieved 18.0 BLEU score and 0.350 chr-F score.

The model in question was trained using Marian-NMT (Junczys-Dowmunt et al., 2018). 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 (Kudo and Richardson, 2018) 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<<).

4.2. Training from scratch

There exist several different frameworks to use with natural language processing tasks, each with their own advantages and shortcomings. One of them is fairseq (Ott et al., 2019) – a sequence modeling toolkit written in PyTorch for training models for translation, summarization, and other 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 specific needs and many implementations for different tasks have been proposed since the fairseq's introduction in 2019. In addition to different architectures, they also provide pretrained models and preprocessed test sets for different tasks, but sadly none of them is in Slovene.

For training our model from scratch, we have decided to use an extension of fairseq (stevezheng23, 2020) that has additional data augmentation methods. We have trained our general model on a corpus described in Subsection 3.1.

4.2.1. 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 Slovene datasets. We first normalized the punctuations, removed non-printing characters, and tokenized both corpora with Moses tokenizer (Koehn et al., 2007). We removed all the sentences that were too short (2 tokens or less) or

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.

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.

too long (250 tokens or more), and the ones where the ratio of lengths was too big because there is a good chance that these kinds of sentences are not translated properly. We then applied Byte pair encoding (BPE) (Sennrich et al., 2016) 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 45,000 types. In the end, we split the data into a training and validation set.

4.2.2. Training

We trained a transformer (Vaswani et al., 2017) model with 5 encoder and 5 decoder layers in the fairseq framework. We used Adam optimizer, an inverse square root learning rate scheduler with an initial learning rate of $7e^{-4}$ and dropout. We also used the proposed augmentation with a cut-off augmentation schema that randomly masks words and this way produces more training data and a more robust translator.

We trained our model for 8 epochs with the mentioned initial learning rate, after which the minimum loss scale (0.0001) was reached, meaning that our loss was probably exploding. We tried training one more epoch with a lower initial learning rate and obtained an even worse performance with the minimum loss scale reached again. That is why we decided to stop the training at 8 epochs. Results of all the epochs are shown in Chapter 5.

4.3. Fine-tuning on TED talks

We preprocessed the TED data in the same way as the general, only this time we used the same dictionary as before and we did not build a new one. Less than 0.1% of tokens in training and validation sets were replaced with *unknown* tokens, so our original dictionary was evidently large enough. We used the best performing epoch from our general translation model (according to the loss on our validation set) for fine-tuning it on our domain data. We trained three different models with three slightly different configurations – one with the same augmentation parameters as the general model, one with increased masking probability and decreased dropout and initial learning rate, and one without augmentation. We trained all of the models for 100 epochs and we are presenting the results of the best epoch for each of them.

4.4. Evaluation

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

The automatic evaluation was carried out on two validation sets. First, the general translation model was evaluated on a subset of the general data, which was split in the preprocessing step (hereinafter referred to as the general validation set). All three models were evaluated on a subset of the domain data (hereinafter referred to as the domain validation set). The manual evaluation was only performed on a subset of the domain validation set, as described in Subsection 4.4.2.

4.4.1. 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 cases like this, automatic evaluation metrics are often used. Natural language is quite subjective. Hence, the perfect measure does not exist, but by evaluating our results with different techniques, we were able to assess the performance of our translation model and compare it with other models. We used automatic metrics most often used in NLP tasks – namely BLEU, chr-F, GLEU, METEOR, NIST, and WER.

4.4.2. Manual evaluation

The translations were also evaluated manually, namely by the fluency-adequacy criterion first described by Church

(Church, 1993). For this part of the evaluation, the Excel format was used. We extracted 6 paragraphs containing 10 consecutive segments from each speech to ensure that the context was clear. Three evaluators (the translators from our group) were assigned 20 segments each. 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 the translation is grammatically well-formed, contains the correct spelling, is intuitively acceptable, and can be sensibly interpreted by a native speaker. To test the adequacy, the evaluator compares both, the source text and the translation, whereas, in the process of the fluency evaluation, the focus is merely on the translation. The evaluators had to provide the scores on a scale from 1 to 4. We chose this evaluation technique because it clearly and simply summarizes and presents the quality of the translations. Since we evaluated three different translation models (pretrained, general, and domain), we had to evaluate the same segments of texts three times. Evaluating one text multiple times by the same person is not recommended, therefore, the translations were exchanged between the three evaluators at the beginning of the evaluation of each translation model.

4.4.3. End user comprehensibility questionnaire

Finally, we evaluated the domain machine-translated texts from the end-user's point of view. Evaluators, who were not familiar with the content of this project, were given the translated texts from the domain model and a questionnaire formed by the translation team of this project. The objective of this questionnaire was to examine whether the end-users understand the information given in the translation, meaning it tested the functionality of the text. The questionnaire was given to nine persons, each evaluating 20 segments from two different speeches - the segments were identical to segments used in the manual evaluation. In the end, we obtained three evaluations for each text (6 speeches altogether). The questionnaire included the following questions:

1. How comprehensible is the text?
2. To what degree does the text seem like it was produced by a native speaker of Slovene?
3. How would you grade the text as a whole?
4. What is the main message of the text?
5. What do you consider as the most problematic part of the text?

For the first and second question, the end-users answered on a scale from 1 to 4, with 1 meaning 'not at all' and 4 meaning 'very much'. The third answer had to be a score from 1 to 4. The fourth question had to be answered with one sentence, and for the fifth question, they had to choose between the following answers: 'unknown words', 'too little context', 'wrong syntax', and 'other'. We chose this evaluation technique because it shows whether the translation is, in fact, functional and useful to the end-user.

5. Results

For the training of our models, we used the Slovenian national supercomputing network that provides access to cluster-based computing capacities. We used the Arnes cluster which is equipped with 48 NVIDIA Tesla V100S PCIe 32GB graphic cards. When training on two of them, one epoch took approximately 4 hours for the general translation model and one minute for fine-tuning on the TED data.

5.1. Automatic evaluation results

In Table 5, we present the quantitative results of the automatic evaluation for the pretrained, general, and domain models.

5.2. Manual evaluation results

Along with the automatic evaluation metrics, we also performed a manual evaluation which provided a valuable human insight into the final product and a better understanding of the typology of the mistakes that occurred in the translations. Each validation set was assessed by two evaluators at all three stages of the model development. The results presented in Table 4 represent the average value of the fluency and adequacy scores for the pretrained, general, and domain models, respectively.

MODEL	Fluency	Adequacy
Pretrained	2.99	3.09
General	2.83	2.9
Domain	2.71	2.9

Table 3: Manual evaluation results on the TED validation set.

5.3. End-user comprehensibility questionnaire results

We received feedback from the end-users based on the questionnaire for the texts from the domain translation model. The average score of the answers that could be interpreted numerically is presented in Table 4. According to the answers to the question 'What is the main message of the text?', the users have, for the most part, understood the text to the degree where they could sufficiently summarize the content. The most frequent answer to the last question (What do you consider as the most problematic part of the text?) was 'wrong syntax', followed by 'lack of context' and 'unknown words'. The participants also pointed out that the general structure of the text was rather confusing.

Text	Question 1	Question 2	Question 3
1	1.33	1	1
2	2	1.33	1.33
3	3	2	2.33
4	1.66	1	1.33
5	2	1.66	1.66
6	2.33	1.66	2
All	2.05	1.44	1.61

Table 4: End-user feedback results from the questionnaire with average scores on a scale from 1 to 4.

Dataset	Metric	Pretrained	General (epochs)								Domain		
			1	2	3	4	5	6	7	8	Configuration 1	Configuration 2	Configuration 3
General	BLEU	-	0.387	0.398	0.405	0.409	0.411	0.417	0.417	**0.420**	-	-	-
	chr-F	-	0.606	0.616	0.619	0.624	0.625	0.629	0.629	**0.629**	-	-	-
	GLEU	-	0.391	0.401	0.407	0.411	0.413	0.417	0.417	**0.420**	-	-	-
	METEOR	-	0.545	0.556	0.560	0.565	0.566	0.569	0.569	**0.571**	-	-	-
	NIST	-	8.752	8.922	8.987	9.063	9.096	9.144	9.114	**9.177**	-	-	-
	WER	-	0.518	0.508	0.503	0.501	0.496	0.497	0.498	**0.494**	-	-	-
Domain	BLEU	**0.192**	0.155	0.167	0.168	0.171	0.175	0.175	0.168	0.179	***0.182***	0.173	0.114
	chr-F	**0.514**	0.487	0.496	0.495	0.497	0.500	0.498	0.500	***0.505***	0.503	0.497	0.440
	GLEU	**0.230**	0.201	0.211	0.212	0.214	0.217	0.218	0.213	0.222	***0.224***	0.216	0.167
	METEOR	0.420	0.398	0.407	0.409	0.409	0.414	0.412	0.416	0.420	**0.426**	0.416	0.346
	NIST	**5.481**	4.877	5.067	5.105	5.132	5.151	5.179	5.074	5.230	***5.344***	5.209	4.228
	WER	**0.659**	0.711	0.696	0.694	0.690	0.689	0.689	0.698	0.685	***0.667***	0.680	0.756

Table 5: Evaluation scores for all models and all validation datasets. The best scores for each dataset and each metric are shown in bold. If the best score was the pretrained model, the second best score is shown in bold and italic to showcase our best score.

6. Discussion

Looking at the results in Table 5, we can first see that on the general validation set, the final epoch of our general model performs the best according to most metrics. This is expected, as the general validation set is comprised of the texts from the corpora that we used for training, so our model may be overfitted on this dataset.

Connected to this, all of the results in the domain validation set are considerably worse than in the general dataset. We can account this to the fact that the domain validation set is truly different from the main training data. As to why the pretrained model in most aspects performs better than our fine-tuned model, we assume that our domain data is not specific enough. Therefore, we could not really fine-tune our model to any specific styles or words, nor were we able to do that in the validation set. The pretrained model performs better because it is trained on a larger dataset than our domain model is fine-tuned on – the TED corpus is relatively small even though we included some additional texts.

Similarly, the results of the manual evaluation showed that the pretrained model produced the most fluent translations with an average score of 2.99 out of 4. This model also achieved the highest score in the adequacy criterion. If we take a closer look at the results of the other two models, it can be seen that both models faced similar difficulties in translating phrasal verbs, terminology, word order, and other lexical structures. The manual evaluation results are relatively low: the general and the domain model received an average of less than 3 points, in both fluency and adequacy. The following examples show the discrepancies between the pretrained model and the other two models on the syntactic, semantic, and morphological levels:

Original: *So then, what is our gut good for?*
Pretrained: *Torej, za kaj je naš občutek dober?*
General: *Torej, kaj je naš črevo dobro za?*
Domain: *Kaj je torej naš črevesje dobro?*

Original: *And I was not only heartbroken, but I was kind of embarrassed that I couldn't rebound from what other people seemed to recover from so regularly.*
Pretrained: *Ne samo, da me je zlomilo srce, ampak me je bilo sram, da se nisem mogel odvrniti od tega, kar so si drugi ljudje zdelo, da si je opomoglo tako redno.*
General: *In nisem bil samo zlom srca, ampak sem bil neprijetno, da se nisem mogel odvrniti od tega, kar se je zdelo, da se drugi ljudje tako redno opomorejo.*
Domain: *In nisem bil le srčni utrip, ampak sem bil neprijetno, da nisem mogel vrniti od tega, kar se je zdelo, da se drugi ljudje tako redno opomorejo.*

However, a quick analysis of the evaluation rates showed that the lowest ratings for the domain model appeared in segments with specialized vocabulary, for example: *"Ampak ko gre za res velike stvari, kot bo naša kariera ali kdo se bo poročil, zakaj bi morali domnevati, da so naše intuicije bolje kalibrirane za te kot počasne, pravilne analize?"* vs the original: *"But when it comes to the really big stuff, like what's our career path going to be or who should we marry, why should we assume that our intuitions are better calibrated for these than slow, proper analysis?"*, and in segments with a higher register, for example, the eloquent text on immigrants: *"Ta vprašanja so protipriseljenska in nativistična v svojem jedru, zgrajena okoli neke vrste hierarhične delitve notranjih in zunanjih oseb, nas in njih, v katerih smo pomembni le in ne." vs the original: "These questions are anti-immigrant and nativist at their core, built around a kind of hierarchical division of insiders and outsiders, us and them, in which only we matter, and they don't."*. In both cases, the rate was never lower than 2.8. The highest rated segments (with the score above 3) included short and simple sentences with everyday vocabulary, such as *"In rekla mi je: Samo dihajte."* or *"Na srečo kriminalci podcenjujejo moč prstnih odtisov."*. Based on the evaluation results, it appears that our domain model would be more valuable in translating general texts with a neutral style and vocabulary.

The group members that evaluated these segments had been participating in this project from the very beginning, so it was crucial to obtain a more objective assessment of our models. Looking at the results from Table 5, the gathered feedback from the questionnaire revealed that overall, the end-users thought that the texts are relatively comprehensible, but are not at all seen as being produced by a native speaker of Slovene. For the first two questions, for which the answers were chosen on a scale from 1–4 (1='not at all'/2='little'/3='good'/4='very much'), only two texts received a score lower than 2 in terms of comprehensibility. When grading the texts, the highest average score for a specific text was 2.33, while the lowest is 1. This variation occurs because not all of the chosen texts were equally complex. For the highest graded text, we received similar responses to the question asking what the main message of the text was: *Opisovanje prstnih odtisov./Puščanje prstnih odtisov./Prstni odtisi poleg vizualne sledi pustijo tudi sled na molekularnem nivoju.* There were only two out of eighteen answers stating that the message was not clear and where the end-users could not summarize the main message, i.e. in texts 1 and 5. The fact that the end-users were in almost all cases able to summarize the main message in one sentence shows that comprehension of the text was still possible despite a large number of significant mistakes (wrong syntax, unknown words, lack of context, changing genders, etc.).

The following examples, segments from text 2, text 3, and text 6, which have also been scored above average in manual evaluation, support this claim:

Original: *And you need something else as well: you have to be willing to let go, to accept that it's over.*
Domain: *Potrebujete tudi nekaj drugega : biti morate pripravljeni pustiti, da sprejmete, da je konec.*

Original: *I'm talking about an entire world of information hiding in a small, often invisible thing.*
Domain: *Govorim o celotnem svetu informacij, ki se skrivajo v majhni, pogosto nevidni stvari.*

Original: *Five years ago, I stood on the TED stage, and I spoke about my work.*
Domain: *Pred petimi leti sem stal na odru TED in govoril o svojem delu.*

Unfortunately, the final version of a machine translator did not meet our expectations regarding the quality of the translations. Some of the major flaws that appeared in the translations were wrong syntax, untranslated words, incomprehensible grammatical structures, wrong use of terminology, and wrong translations of polysemes. While we expected the machine translator to be inappropriate for translating complex sentences, we were surprised that it did not perform well when translating even basic grammatical structures. Here are two examples:

Original: *So then, what is our gut good for?*
Domain: *Kaj je torej naš črevesje dobro?*

Original: *I later found out that when the gate was opened on a garden, a wild stag stampeded along the path and ran straight into me.*
Domain: *Kasneje sem ugotovil, da ko so vrata odprta na vrtu, je divji stag žigosanih po poti in tekel naravnost v mene.*

Original: *And for two years, we tried to sort ourselves out, and then for five and on and off for 10.*
Domain: *Dve leti smo se poskušali razvrstiti, nato pa pet let in več.*

The reasons for the poor functioning of the machine translations could be numerous. It is possible that we have not collected enough data or that the chosen data might not have been the most suitable for this project. We estimate that the main factor that impacted the final results the most is the wide range of different topics covered in TED Talks. This means that our domain translation model did not focus on just one domain and, essentially, there was not enough specific data from which it could train. What is more, the initial data consisted of transcriptions of English spoken discourse and their Slovene translations in the form of subtitles. It is important to keep in mind that neither spoken discourse nor subtitles have characteristics typical for standard text types. Finally, not all of the chosen texts were equally complex and they had different syntactic, morphological, and lexical features. Therefore, some of the texts in the data were essentially too difficult to translate.

7. Conclusion

The main purpose of this project was to develop a tool that would automatically provide Slovene transcriptions or subtitles for English TED Talks. Our domain translation model provides translations that convey the main message of the texts, is based on the appropriate methodology, and built with all the necessary tools. Even more, the results of automatic metrics showed that it is comparable to other neural machine translation models. On the other hand, the lack of a uniform training dataset resulted in poor and incomprehensible translations. However, we believe that acknowledging all of the discussed shortcomings in future research could significantly improve the development of speech-to-text and translation technologies for Slovene language users. Neural machine translation is still relatively new and will develop in the following years because it is useful for translators and the general public. Our project contributed to the advancement of the field and could provide valuable information for similar work in the future.

Acknowledgments

We would like to thank our mentors, Slavko Žitnik, Špela Vintar, and Mojca Brglez, for helping us with the project. We would also like to thank the nine evaluators who provided end-user feedback by filling out our questionnaire.

We would also like to thank SLING for giving us access to powerful graphic cards to successfully finish our training, as we would still be training our general model without them. Special thanks to Barbara Krašovec from Arnes support who helped us with our numerous problems when trying to connect to their cluster.

8. References

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Comput. Linguist.*, 16(2):79–85.

Kenneth Church. 1993. Good applications for crummy machine translation. *Machine Translation*, 8:239–258, 12.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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. 2018. Marian: Fast neural machine translation in C++. In: *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. 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*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2021. Pre-training multilingual neural machine translation by leveraging alignment information.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

stevezheng23. 2020. fairseq_extension.

Jörg Tiedemann. 2012. Parallel data, tools and inter- faces in opus. In: *Proceedings of the Eight Interna- tional Conference on Language Resources and Evalua- tion (LREC'12)*, Istanbul, Turkey, may. European Lan- guage Resources Association (ELRA).

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In: *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online, November. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.