FRI Natural language processing course 2021
Project report

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

Automatic language translation (joint work with Faculty of Arts)

Marjan Bogataj, David Peklenik Urbanč, Aljaž Dobnik, Ilda Šećkanović, Saša Grahovac Fabbri, Jan Gajski

Abstract

This article presents a joint project between students from the Faculty of Computer and Information Science and Faculty of Arts. The main idea was to establish a natural language processing model for translating from English to Slovene and vice-versa. To complete this task, OpenNMT-py neural translation framework was selected and used with parallel general corpora for the model to learn basic language translations. Afterwards, the model was further trained on smaller corpora containing medical and pharmaceutical texts in an attempt to improve the model in a specific vocabulary setting. The training was executed using Google Colab. The translation models were created with different parameters in order to evaluate their influence on translation quality. A small sample of translated texts was manually evaluated using the adequacy-fluency metric. The same sample was further assessed with the BLEU method. The results show a large discrepancy between machine translated sample and human reference translations, with many inadequacies and elements that make the machine translated texts sound awkward and artificial, difficult to read and comprehend. To improve the quality of machine translation, future research could include verifying the quality of the training corpus, using a larger base vocabulary, and training and validating the model using simpler sentences containing less field specific terminology.

Keywords

natural language processing, neural machine translation framework, general corpora, specialized corpora, model training, fluency, adequacy, BLEU

Advisors: Slavko Žitnik, Mojca Brglez, Špela Vintar

Introduction

We begin by referencing related work in the field of language processing and translation with focus on evaluation and comparison of human and machine translation. We then describe the data and methods used in the project. This is followed by the result and discussion sections, where we evaluate the process of preparing and training the machine translator, and assess the machine translation and evaluation results. We conclude the article with a brief summary and suggest potential improvements of our research process.

Related work

Due to the increasing use of machine translators in the translation process, there is more and more research dedicated to evaluating the quality of machine translation (MT). Evaluation of MT can be done either automatically, where computer metrics such as BLEU, GTM, METEOR compare MT with human translation, or manually, where a variety of assessment methods are used to evaluate adequacy, fluency, terminology, certain types of errors, or compare more translations of one text (Kuzman 2019 [1], 23–27) ...

Bordon (2021) [2] analysed the end-user comprehensibility of unedited machine-translated texts. In his questionnaire, he included texts translated by Google Translate and eTranslation from English into Slovene. Participants did not have access to the original text and saw translations in two types of contexts – either with an image or without. The results of the survey showed that the average comprehension rate was at 59%, with eTranslation achieving better results than Google Translate.

Vintar (2018) [3] dealt with the evaluation and comparison of a statistical phrase-based and neural model of Google Translate. She compared the translations between English and

Slovene and focused on karst terminology. Overall, the analysis showed that the neural translator achieved better results in both directions, but when translating terminology, it was better only in the direction from English to Slovene – from Slovene to English, the statistical model was more successful.

Kuzman (2019) [1] compared Google Translate and two of her models, which were specialized in translating literary texts from English into Slovene. The results of automatic and manual evaluation methods, except the fluency method, showed that a model tailored to a specific author was more successful than a model that was trained on a bigger corpus of different literary works. However, the best results with all evaluation methods were achieved by Google Translate.

Data

The basic translation model was trained using the TC3 parallel corpus. It is a collection of multiple Slovene-English corpora: OpenSubtitles 2018, Europarl, EMEA, DGT, and ELRC. The combined corpus consists of 24,419,756 sentences which further contain 850,245 English and 1,369,057 Slovene unique words. TC3 corpus was downloaded in sentence aligned format (i. e., two monolingual text files, where each line represents one sentence), which had special characters separated from words with the space character, which is important for building the vocabulary and the training of the model. It had some characters replaced with a flag as well (for example single ' and double quotation marks " were replaced with *'* and *"*). TC3 corpus was then randomly shuffled and segmented into two separate parts:

- Training data - used for training the basic model with the translation framework, contains most of the corpus data
- Validation data - used for the evaluation of a model in the process of training, it is used to correct the model, contains about 5000 sentences, as it is recommended on OpenNMT-py's Github page [4]

Usually, the corpus would be segmented into the test data used for the model evaluation, but the test corpus was already pre-determined by the assistant. It contains around 2000 sentences, which had to be converted into the same format as the training corpus (a space was added before and after all special characters if it was not already present).

In the experimental phase, our models were trained exclusively using the Europarl parallel corpus, which is also present in TC3. The corpus is a collection of parallel texts in 21 European languages from the proceedings of the European Parliament, dating from between 1996 and 2011 (Koehn 2005) [14]. For this project, we used only the Slovene-English parallel subcorpus. It consists of 623,490 sentences which further contain around 13 million Slovene and over 15 million English words.

To specialize the model to the field of medicine, we combined two separate corpora : ECDC (European Centre for Disease Prevention and Control) and EMA (European Medicines Agency), totaling about 340,000 sentences, 31,716 English and 36,566 unique Slovene words, which were not present in the general corpus TC3. Both EMA and ECDC were converted first to the same format - two monolingual text files, where each line represents one sentence, then we combined them, randomly shuffled, added a space before and after all special characters if it was not already present (for example ",.-*?!'&), removed any potential empty sentences and finally, it was split into train, validation (around 5000 sentences) and test (around 2000 sentences) subsets.

Methods

The framework chosen for this task was OpenNMT-py [4], developed by the University of Harvard and SYSTRAN and is currently maintained by SYSTRAN and Ubiqus. OpenNMT-py is an open-source neural machine translation system designed to be research-friendly and can be used for natural language translation. It also uses PyTorch, which is widely used across machine translation frameworks. OpenNMT-py was chosen for its ease of use and well-written documentation. The project was executed in three separate steps:

- The first step was the preparation, which started by combining multiple corpora into one. Then it was randomly shuffled, converted into the correct format, cleaned, corrected, and finally, split into train, validate and test. This was done on a local PC (Personal Computer) using Python 3.7. We selected an appropriate folder structure and uploaded the project to GitHub.
- The second step was the training of the model, which was executed using Google Colab due to its processing resources. First, a vocabulary was built using both the TC3 and the specialized corpora, then both the English-Slovene and Slovene English models were trained, first only on the TC3 corpus and then additionally on the specialized corpus for medicine. Finally, the models were used for the translation of both the test set of the general and the specialized corpora. The general model was also used to translate the test set for the specialized corpus to see the difference between the general and specialized models. The English-Slovene model was also trained with a smaller vocabulary size but longer training time.
- The third and final step was the evaluation of all translations, which was executed again using Python, on a local machine.

Evaluation methods

Despite very accurate machine evaluation of machine translated texts, human judgment is essential for designing effective evaluation systems and interpreting the scores they provide. Human input is crucial when it comes to improving MT evaluation systems since human analysis often serves as a framework for the creation of such tools.

In the present day, the most commonly used method, developed by TAUS, is used to evaluate neural-machine translated texts through adequacy and fluency. The TAUS method requires evaluators to read and evaluate the same text. At this stage, we have decided to evaluate our texts via adequacy and fluency, where both adequacy and fluency are evaluated on the scale from 1 to 4. When assessing fluency, the lowest value stands for the text being incomprehensible (1) and the highest (4) for the text being flawless. When assessing adequacy, the lowest value (1) means that the original meaning is not perceived at all from the translation and the highest value (4) means that the meaning is fully perceived from the machine translated text.

Each of the reviewers evaluated 90 sentences, whereby sentences, consisting only of numbers or other symbols (for example: 12.5mg or EU/1/99/113/002), were left out to not influence the final rating. We put all our ratings into an excel spreadsheet and then calculated the average rate that was used for a comparison between the adequacy and the value obtained by the BLEU metric method. BLEU scores can range between 0 or 0% (completely different from the reference translation) and 1 or 100% (same as the reference translation) (Shterionov et al. 2018). Our evaluation score was computed at a document level and it was calculated on the free Interactive BLEU score evaluator on the Tilde website [5].

We hope to show how the most common inadequacies, realized by the neural machine translation model, can be addressed and how this can contribute to developing a new and better system for neural machine translation. Our findings will be presented in the Evaluation Results section.

Algorithms

Firstly, the specialized corpora by ECDC and EMA were combined into one single corpus using *combine_corpus.py*. Then, all lines were randomly shuffled in the same order for both the Slovene and English corpus with *randomize_corpus.py*. The newly made specialized shuffled corpus alongside the test set for the general model were converted into the correct format, required by OpenNMT-py, where special characters are separated from everything with a space. This was done using the *convert_korpus.py* Python script. The specialized corpus was then split into train, validate and test subsets and the general corpus TC3 was only split into train and validate, since the test set was provided separately.

The vocabulary was built on Google Colab using both the specialized and the TC3 corpora from over 24 million sentences, using the *onmt_build_vocab* OpenNMT-py command. The process took five minutes and resulted in over 800,000 English and 1.4 million Slovene words. Next, the general translation model was trained using both the vocabulary and the train subset of the TC3 general corpus. As this command was executed on Google Colab, there were system memory limitations, whereby the model could not be trained using the full length of the vocabulary. There was also only one GPU card available, so the process was slow. The graphics card used for training the model was Tesla T4.

The parameters for training the model are the location of English and Slovene vocabularies, the location of the *train* and *validation* files for both languages, number of GPUs to train on, location of the saved model, number of steps (iterations) the model will train for, and how often the model will be evaluated. These parameters are stored in *conf.yaml* file, which is used with OpenNMT-py command *onmt_train* to train the model.

The first model was built using only 50,000 most common words from the vocabulary, with validation on every 5000th step. The general model was trained on the TC3 train corpus with 100,000 steps, with the process taking over three hours. It was then specialized with the specialized corpus and additional 30,000 steps, which took one and a half hours. Since the training process took multiple hours, it led to some problems with Google Colab, and the session would end prematurely if it was not being interacted with or if it exceeded the time limit of 12 hours per day. Both of these scenarios would lead to a complete loss of progress if it was not saved in between.

This is why all subsequent models were trained with a low number of steps (10,000 for the general and 5,000 for the specialized model), but with a large vocabulary size of 100,000 most common words. This led to doubling the size of the previous model, since it had to learn twice the amount of words. Training the model with a vocabulary size larger than 100,000 resulted in a memory error, which is why the model uses less than 15% of the created vocabulary. The training times of these models are shown in the table below.

Model type	EN-SLO	SLO-EN
General	18 min	1h 10 min
Specialized	12 min	1h 20 min

Table 1. Training time of different models

Finally, the trained models can be used to translate from one language to the other. This is done with the OpenNMT-py command *onmt_translate* and the pre-generated *test* corpus. The translation is also being evaluated at the same time with a built-in metric called PRED AVG SCORE, which represents the average logarithmic confidence for word generation. The higher scores signify better translations. The scores are shown in Table 2.

Type	EN-SLO	SLO-EN
General model and test	-1.5518	-1.4068
Specialized model and test	-0.5991	-0.6473
General model and spec. test	-1.4611	-1.4844

Table 2. Average translation prediction score

The translated texts were compared with their corresponding language test corpus in order to determine its quality. Multiple metrics were used for this purpose: BLEU, CHRF, GLEU, METEOR, NIST, and RIBES.

Model	Language	BLEU	CHRF	GLEU	NIST	RIBES	METEOR
General - 10k steps, 100k vocab	EN-SL	0.0421	0.1893	0.0625	1.5254	null	0.2113
	SL-EN	0.0325	0.1800	0.0502	1.2350	null	0.1899
Specialized - 15k steps, 100k vocab	EN-SL	0.2198	0.3934	0.2061	4.9317	null	0.4096
	SL-EN	0.2873	0.4816	0.2887	6.5467	null	0.5171
General - 100k steps, 50k vocab	EN-SL	0.1756	0.3593	0.1968	4.5171	null	0.3829
	SL-EN	/	/	/	/	/	/
Specialized - 130k steps, 50k vocab	EN-SL	0.2422	0.4194	0.2571	5.6063	null	0.4454
	SL-EN	/	/	/	/	/	/

Note: **k** in "General - 10k steps, 100k vocab", for example, signifies thousands

Table 3. Metric scores for specialized test set

Model	Language	BLEU	CHRF	GLEU	NIST	RIBES	METEOR
General - 10k steps, 100k vocab	EN-SL	0.0513	0.2357	0.0813	1.9600	null	0.2455
	SL-EN	0.0536	0.2492	0.0815	1.9861	null	0.2546
General - 100k steps, 50k vocab	EN-SL	0.1850	0.3917	0.2152	5.2656	null	0.4055
	SL-EN	/	/	/	/	/	/

Note: **k** in "General - 10k steps, 100k vocab", for example, signifies thousands

Table 4. Metric scores for general test set

- BLEU (bilingual evaluation understudy) claims to measure the quality of the translation similar to a human [6]
- CHRF (Character n-gram F-score) uses a character n-gram F-score for automatic evaluation of machine translation output [7]
- GLEU (Google-BLEU) - used for measuring sentence level similarity [8]
- NIST (National Institute of Standards and Technology) is based on BLEU, with added weighing to n-grams [9]
- RIBES (Rank-based Intuitive Bilingual Evaluation Score) is an automatic evaluation metric for machine translation [10]
- METEOR (Metric for Evaluation of Translation with Explicit ORdering), which is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. [11]

These metrics were calculated in *evaluate_result.py* for eight different combinations of models and test inputs, which are shown in tables 3 and 4. Unfortunately, there was a problem with computing RIBES, which is why it is listed as "null" in both tables.

Results

As established, the model was trained in two different configurations - with different number of training steps (a setting, which directly impacts how long the training will take), vocabulary size and validation frequency.

In the first scenario, the English-Slovene model was trained with 130,000 steps and evaluated every 5000th step (meaning 26 times for this scenario). It was first trained on the general corpus with 100,000 steps and then on the specialized corpus for additional 30,000. Example 1:
SENT 1820: ['There', 'is', 'no', 'experience', 'regarding', 'the', 'administration', 'of', 'Irbesartan', 'Krka', 'in', 'patients', 'with', 'a', 'recent', 'kidney', 'transplantation', '.']
PRED 1820: Pri bolnikih z nedavno presaditev ledvic ni izkušenj z zdravilom Irbesartan Krka .
Example 2:
SENT 1854: ['These', 'two', 'drugs', 'lower', 'the', 'pressure', 'in', 'the', 'eye', 'by', 'different', 'mechanisms', '.']
PRED 1854: Ti dve $<$ *unk* $>$ $<$ *unk* $>$ tlak v očesu z različnimi mehanizmi .

The second scenario had a bigger vocabulary size of 100,000 words, lower number of training steps of 15,000 and higher validation frequency of 2500 steps. Example 1:
SENT 1820: ['There', 'is', 'no', 'experience', 'regarding', 'the', 'administration', 'of', 'Irbesartan', 'Krka', 'in', 'patients', 'with', 'a', 'recent', 'kidney', 'transplantation', '.']
PRED 1820: O uporabi zdravila Irbesartan Krka pri bolnikih z nedavno ledvično insuficienco ni izkušenj .
Example 2:
SENT 1854: ['These', 'two', 'drugs', 'lower', 'the', 'pressure', 'in', 'the', 'eye', 'by', 'different', 'mechanisms', '.']
PRED 1854: Ti dve zdravili , ki so odvisni od krvnega tlaka v očesu .

Evaluation results

By comparing examples for both models, we can conclude that the first model with a smaller vocabulary knows the sentence structure better and if the words in the sentence are not too specific, it translates the sentence well. However, the smaller vocabulary size means that some basic words like "drugs" are not translated, resulting in an unfinished sentence . By looking at Table 3, we can compare both models, and it is evident that the model with a smaller vocabulary but bigger number of steps is better than the other one, since it has a better score at every single metric.

Human evaluation

Each segment was graded on a scale of 1 to 4, which was then changed to a scale from 0% to 100%. Table 5 shows the average scores of all segments for each individual and all evaluators together. However, due to the nature of the project, we could see the original when assessing fluency.

Fluency	Rater 1	27%
	Rater 2	28%
	Rater 3	30%
	Together	28.5%
Adequacy	Rater 1	29%
	Rater 2	31%
	Rater 3	37%
	Together	32.5%

Table 5. Average scores for fluency and adequacy

Based on the assessment of fluency, we can say that the evaluators had very similar ratings and there were no major discrepancies between them. Slightly larger differences occur in adequacy. This is because the instruction for fluency is more specific and guides us on what exactly we need to pay attention to. Meanwhile, the instruction for adequacy is very broad, so each individual understands the instruction differently and subjectively evaluates the translation. Nevertheless, the translation was relatively poorly graded due to errors in the translation.

ERRORS

The human evaluation of machine translated documents showed many inadequacies and difficulties that can be mitigated with further research and training of the neural translation models. The biggest problem we encountered with MT is that many sentences and parts of the sentence were left untranslated and marked as $< unk >$:

Patient monitoring on day 1 should include vital signs including pulse, blood pressure and respiratory rate. $\rightarrow <$ *unk* $> <$ *unk* $> <$ *unk* $> <$ *unk* $> <$ *unk* $> <$ *unk* $> <$ *unk* $>$.

Other errors:

- untranslated terminology *Phospholipidosis in pulmonary macrophages due to accumulation of memantine in lysosomes was observed in rodents.* $\rightarrow <$ *unk* $>$ *pri* $<$ *unk* $> <$ *unk* $> <$ *unk* $>$ *so opazili pri glodalcih.*
- repetition *If you experience prolonged bleeding when taking Plavix. If you cut or injure yourself, it may take longer than usual for bleeding to stop.* $\rightarrow$ *Če ste vzeli zdravilo Plavix Če ste vzeli zdravilo Plavix Če ste vzeli zdravilo Plavix Če ste prenehali jemati zdravilo Plavix Če ste vzeli ali ste vzeli zdravilo Plavix, lahko vzamete dlje kot običajno za krvavitev.*
- personal and possessive pronouns used in the same sentence that resulted in the wrong translation *You should inform your doctor* $\rightarrow$ *Zdravniku morate obvestiti svojega zdravnika*

Interestingly, the translator did not invent new words, as is usual for NMT. On the other hand, as already mentioned, many words remained untranslated, which is visually immediately noticeable and consequently degrades the quality of the translation.

Automatic evaluation

According to many researchers, BLEU scores are very similar to human evaluation, and therefore, it is also widely used [12]. However, researchers ([13], [3]) show that BLEU is, in many cases, better at evaluating PBSMT (phrase-based statistical MT) rather than NMT. That means that automatic evaluations do not necessarily always conform with the quality of the NMT.

To see if BLEU will evaluate similarly to us, we used the average value of adequacy to compare it with BLEU, as both methods compare MT and reference translation. This evaluation score was computed at a document level, so both files – human translated file and MT file – included 90 segments. From Table 6, we can see that the BLEU metric evaluates

Adequacy	BLEU
32.5%	24.23%

Table 6. The total average value of adequacy and BLEU score

translation worse than we have, with a difference of 8.27. In any case, percentages are low, which means that the translated texts are difficult to read and are quite different from the reference translation.

Experiments

The model was trained on the Europarl corpus [14] using different settings in order to explore which parameters impact the model and how. Three different settings for the number of steps were used and two different frequencies of model evaluation during training. This was done for both translation models (English to Slovene and vice-versa).

In the first scenario, the English-Slovene model was trained with only 1000 steps and evaluated every 500th step (meaning twice for this scenario). The results were poor since most of the sentences for translation comprised of unknown words

(tagged as $< unk >$). Most of the translations proved to be very similar and contained many repetitions. Example of the first scenario:
SENT 6009: ['This', 'is', 'the', 'proof', 'that', 'fiscal', 'stability', 'leads', 'to', 'growth', 'and', 'employment', '.']
PRED 6009: To je zelo pomembno , da je $< unk >$.
PRED SCORE: -15.0615

Note: The sentence "To je zelo pomembno" was being repeated in many other translations.

The second scenario had the same validation frequency (every 500 iterations), but the number of steps for training was increased to 3000. This yielded much better results since there were fewer $< unk >$ words, with actual sentences being formed. However, the translations themselves were deficient as the sentences did not make much sense. It seemed as if the model was purely guessing the correlations between both languages, although some words were translated correctly. Example of the second scenario:
SENT 6009: ['This', 'is', 'the', 'proof', 'that', 'fiscal', 'stability', 'leads', 'to', 'growth', 'and', 'employment', '.']
PRED 6009: To je tisto , ki je $< unk >$, da je $< unk >$ $< unk >$ in $< unk >$.
PRED SCORE: -19.7641

In the third scenario, we further increased the number of steps to 6000. This improved the readability of the sentences and greatly improved the accuracy of translations. The unknown $< unk >$ words appeared far less frequently, they were not present in every translation anymore, and some short sentences were even translated correctly. Example of the third scenario:
SENT 6009: ['This', 'is', 'the', 'proof', 'that', 'fiscal', 'stability', 'leads', 'to', 'growth', 'and', 'employment', '.']
PRED 6009: To je dokaz , ki je $< unk >$ $< unk >$ $< unk >$ za rast in zaposlovanje .
PRED SCORE: -7.8747

The last test was done with the same number of steps as the third experiment (6000), but with the increased frequency of model validation (from every 500 iterations to every 250 iterations). This did not change the resulting translations by much, although surprisingly, the average predicted score even worsened, which may indicate overfitting to the validation set.

Example of the last test:
SENT 6009: ['This', 'is', 'the', 'proof', 'that', 'fiscal', 'stability', 'leads', 'to', 'growth', 'and', 'employment', '.']
PRED 6009: To je dokaz , da je za rast in zaposlovanje .
PRED SCORE: -7.6077

The Slovene–English model behaved in a similar way, although there were fewer $< unk >$ words when the number of training steps was 1000. There were a lot of repetitions and the sentences were not understandable. Example:
SENT 5336: ['Ljudi', ',', 'ki', 'bodo', 'skrbeli', 'za', 'kmetijski', 'sektor', ',', 'pa', 'tudi', 'za', 'področja', 'pred', 'in', 'za', 'njim', '.']
PRED 5336: I would like to thank the Commissioner , I would like to thank the Commissioner , I would like to thank the Commissioner .
PRED SCORE: -30.2082

The increase in the number of steps for training to 3000 had a positive impact on the quality of translations, similarly to the English–Slovene model, meaning the translations contained some correct words but the model could not correctly form a sentence. Example:
SENT 5336: ['Ljudi', ',', 'ki', 'bodo', 'skrbeli', 'za', 'kmetijski', 'sektor', ',', 'pa', 'tudi', 'za', 'področja', 'pred', 'in', 'za', 'njim', '.']
PRED 5336: People who will have a $< unk >$ system , but we also have to ensure that we are in favour of a system and for you .
PRED SCORE: -34.6084

Similarly to the English–Slovene model, the increase of steps to 6000 greatly improved the readability of the translations, the number of $< unk >$ decreased, and some shorter sentences were translated correctly. Example:
SENT 5336: ['Ljudi', ',', 'ki', 'bodo', 'skrbeli', 'za', 'kmetijski', 'sektor', ',', 'pa', 'tudi', 'za', 'področja', 'pred', 'in', 'za', 'njim', '.']
PRED 5336: People who are going to be $< unk >$ for agricultural sector , and also for them .
PRED SCORE: -13.7145

Evidently, the longer the model is trained, the better it understands the correlations between languages and therefore produces improved translations. It is also important to note that the vocabulary was built on only the first 10,000 sentences from a 600,000 sentence corpus, but the number of unknown words still decreased rapidly after increasing the number of training steps for the model from 1000 to 3000. The evaluation frequency is not as important, however, it can lead to overfitting of the validation corpus and thus worsening the translation accuracy. This can be observed with the average prediction score, where the example with fewer validations (Example 3) had a score of -0.8293, and Example 4 with doubled validation rate had a score of -0.9301 . There was little difference between the training process between the English–Slovene and the Slovene–English model. They both showed similar results relating to the number of training steps.

Discussion

A number of models were created with different parameters, all of which resulted in slightly or considerably different results. They differed in the number of steps used for training of the model, their validation frequency, vocabulary size and the data on which the model was trained.

The smallest impact of these had the validation frequency,

which can lead to small margins due to overfitting. The vocabulary size, used for the training of the model had a significant impact on the size, accuracy, legibility of the model, but the biggest difference was made by the number of steps taken for the training of the model. We started with 1000 steps and noticed immense difference on every couple thousand steps. The model trained with only 1000 steps could not translate a single word correctly or even knew what a sentence was supposed to look like, whereas the model trained with 3000 steps knew most of that already and model trained with 6000 steps already translated some whole sentences correctly. After that, with the increasing number of steps the model better understood the structure of the sentences and language as a whole. There was little difference between models trained with 10,000 and 15,000 steps, however, there is a big difference between models trained with 15,000 steps and 100,000 steps, as evident in Table 4.

Vocabulary size also matters, but not nearly as much. It helps with sentence formation, because with a larger vocabulary, there are fewer $< unk >$ words present. But if the number of training steps is too low, even though it knows how to translate more words, it cannot use them correctly in a sentence. Since OpenNMT-py trains the model by storing many items in the memory, including the vocabulary, its size can quickly become limited, whereas the number of training steps is limited only by time. Ideally, the model would include the whole vocabulary and was trained with sufficient number of steps that further training would not make a difference.

The general models were also used to translate specialized pharmaceutical data on which it was not trained. This resulted in significantly worse results compared to the translations of specialized models with the same vocabulary size (see Table 3). However, it is worth noting that the general model with 100,000 training steps performed more than four times better than the one with 10,000 training steps, according to the BLEU score. That means that with more training steps, the model can learn to better translate sentences with words on which it was never trained.

Something we did not entirely explore was the size and the quality of the training corpus, which can determine the quality of the translations. If a certain corpus is very big, but has a lot of sentences that do not make sense or the translations are poor in the original corpus, the model could produce worse results compared to one built with a smaller corpus with more quality translations and little corrupt/bad data. The concept of GIGO (Garbage in, garbage out) applies [15].

During manual evaluation of translations we observed that the model had less difficulties with simple, short phrases and sentences e.g. ”6.2 Incompatibilities” that was correctly translated into ”6.2 Inkompatibilnosti”, but even in such sentences it got easily confused if special punctuation marks were included, e.g.”• Grade 3” was translated into ”• $< unk >$ 3”.

By comparing adequacy and BLEU metrics, it can be observed that automatic evaluation does not produce the same results as human evaluation, at least when evaluating NMT. It would be interesting to make a PBSMT in addition to the NMT, compare their BLEU scores, and observe how much human evaluation would differ from these assessments.

Conclusion

Throughout the project, we learned that MT models require adequate fine-tuning and evaluation before they can be considered usable. Various types of errors, as well as the inability of making correlations between languages are a testament to that. We identified some of the most common errors and calculated the average grade for each individual evaluator as well as the BLEU rating for the whole sample text. Considering that the BLEU grade was lower than ours, we can assume that the human evaluation method focuses on different aspects than BLEU. We can assume that the instructions, determined for the human adequacy evaluation are too vague to produce entirely objective results. In any case, both scores were quite significantly low. Based on the research of our colleagues from the Faculty of Computer and Information Sciences and their training of the models, as well as on the research of our colleagues from the Faculty of Arts, and their evaluation, future research should focus on training models with a narrower, more specific vocabulary, containing typical phrases and sentences that will help the model recognize the language specific correlations.

References

[1] Kuzman, Taja. 2019. ≫Nevronsko strojno prevajanje literarnih besedil iz angleščine v slovenščino≪ [Neural Machine Translation of Literary Texts from English to Slovene]. Master’s thesis. University of Ljubljana.

[2] Bordon, David. 2021. ≫Razumevati nevronščino: kako si ljudje razlagamo jezik strojnih prevajalnikov≪ [Comprehending the neural language: How people understand the language of machine translation engines]. Master’s thesis. University of Ljubljana.

[3] Vintar, Špela. 2018. ≫Terminology Translation Accuracy in Statistical versus Neural MT: An Evaluation for the English-Slovene Language Pair≪. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). http://lrec-conf.org/workshops/lrec2018/W19/pdf/7_W19.pdf.

[4] OpenNMT-py, https://github.com/OpenNMT/OpenNMT-py/. Last accessed 1 May 2021

[5] Tilde, https://www.letsmt.eu/Bleu.aspx. Last accessed 21 May 2021

[6] BLEU, https://en.wikipedia.org/wiki/BLEU. Last accessed 20 May 2021

[7] Popovič, Maja. 2015. ≫chrF: character n-gram F-score for automatic MT evaluation≪, Proceedings of the Tenth Workshop on Statistical Machine Translation (392–395), https://www.aclweb.org/anthology/W15-3049

[8] GLEU, https://colab.research.google.com/github/gcunhase/NLPMetrics/blob/master/notebooks/gleu.ipynb. Last accessed 20 May 2021

[9] NIST, https://en.wikipedia.org/wiki/NIST_(metric). Last accessed 20 May 2021

[10] RIBES, http://www.kecl.ntt.co.jp/icl/lirg/ribes/. Last accessed 20 May 2021

[11] METEOR, https://en.wikipedia.org/wiki/METEOR. Last accessed 20 May 2021

[12] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pages 311–318.

[13] Shterionov, D., Superbo, R., Nagle, P., Casanellas, L., O'Dowd, T., & Way, A. (2018). Human versus automatic quality evaluation of NMT and PBSMT. Machine Translation, 1–19.

[14] Koehn, Philipp. 2005. ≫Europarl: A Parallel Corpus for Statistical Machine Translation.≪ MT Summit 2005. https://www.statmt.org/europarl/.

[15] GIGO, https://en.wikipedia.org/wiki/Garbage_in,_garbage_out. Last accessed 21 May 2021

[16] Readability Evaluation Guidelines, published by TAUS, 2017, available via: https://cdn2.hubspot.net/hubfs/2734675/Readability%20Evaluation%20Guidelines.pdf

[17] Escribe, Marie. 2016: Human Evaluation of Neural Machine Translation: The Case of Deep Learning, Guildhall School of Business and Law London Metropolitan University. London. Available via: https://www.aclweb.org/anthology/W19-8705.pdf