# Preparing a Corpus and a Question Answering System for Slovene

**Matjaž Zupanič*, Maj Zirkelbach*, Uroš Šmajdek*, Meta Jazbinšek†**

*Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113, SI-1000 Ljubljana
{mz4689, mz5153, us6796}@student.uni-lj.si

†Department of Translation Studies, Faculty of Arts, University of Ljubljana
Aškerčeva cesta 2, SI-1000 Ljubljana
mj6953@student.uni-lj.si

**Abstract**

Lack of proper training data is one of the key issues when developing natural language processing models based on less-resourced languages, such as Slovene. In this paper we discuss machine translation as a solution to this issue, with the focus on question answering (QA). We use the SQuAD 2.0 dataset, which we have translated using eTranslation machine translator. To improve the reliability of translations, we translate the answers together with the context instead of separately, reducing the rate at which answers were not found in the context from 56% to 7%. For comparison, we also perform manual post-editing of the small subset of machine translations. We then compare these datasets utilizing various transformer-based QA models and observe the differences between the datasets and different model configurations. The results have shown little distinction between monolingual and larger multilingual models: monolingual SloBERTa scored 64.9% exact matches on the machine translated dataset and 72.6% exact matches on human translated one, whereas multilingual RemBERT scored 64.2% exact matches on the machine translated dataset and 71.9% exact matches on human translated one. Additionally, using machine translated dataset in the evaluation produces notably worse results then the human translated dataset. Qualitative analysis of the translations has shown that mistakes often occur when the sentences are longer and have more complicated syntax.

## 1. Introduction

One of the goals in artificial intelligence is to build intelligent systems that would be able to interact with humans and help them. One of such tasks is reading the web and then answer complex questions about any topic over given content. These question-answering (QA) systems could have a big impact on the way that we access information. Furthermore, open-domain question answering is a benchmark task in the development of Artificial Intelligence, since understanding text and being able to answer questions about it is something that we generally associate with intelligence.

Recently, pre-trained Contextual Embeddings (PCE) models like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) and A Lite BERT (ALBERT) (Lan et al., 2020) have attracted lots of attention due to their great performance in a wide range of NLP tasks.

Multilingual question answering tasks typically assume that answers exist in the same language as the question. Yet in practice, many languages face both information scarcity—where languages have few reference articles—and information asymmetry—where questions reference concepts from other cultures. Due to the sizes of modern corpora, performing human translations is generally infeasible, therefore we often employ machine translations instead. Machine translation however is for the most part incapable of interpreting nuances of specific languages such as culturally specific vocabulary or for example the use of articles, indication of grammatical number or gender and conjugation endings when comparing English and Slovene.

In this work we present a method for a construct of a machine translated dataset from SQuAD 2.0 (Rajpurkar et al., 2018) and evaluate its quality using various modern QA models. Additionally, we benchmark its effectiveness by performing manual post-editing on a subset of the translated dataset and comparing the results.

The main contributions of our work are:

- a pipeline for translation of English question answering dataset;
- a Slovene monolingual model SloBERTa, fine-tuned on machine translated data and three different fine-tuned multilingual QA models, M-BERT, XLM-R and CroSloEngual BERT, all on machine translated and both original and machine translated data; and
- comparison of human and machine translated data in terms of question answering performance.

In Section 2 we present the related work. In Section 3 we present our dataset, the process of translation and post-edition, and evaluate the quality of the translation. In Section 4 we give a brief overview of the models used in the evaluation. In Section 5 we present the evaluation and discuss the results in Section 6. In Section 7 we present the conclusions and give possible extensions and enhancements for future work.

## 2. Related work

Early question answering systems, such as LUNAR (Woods and WA, 1977), date back to the 60's and the 70's. They were characterised by a core database and a set of rules, both handwritten by experts of the chosen domain. Over time, with the development of large online text

repositories and increasing computer performance, the focus shifted from such rule-based system to using machine learning and statistical approaches, like Bayesian classifiers and Support Vector Machines. An example of this kind of system that was able to perform question answering on Slovene language was presented by Čeh et al. (Čeh and Ojsteršek, 2009) in 2009.

Another major revolution in the field of question answering and natural language processing in general was the advent of deep learning approaches and self-attention. One of the most popular approaches of this kind is BERT (Devlin et al., 2018), a transformer model introduced in 2019. Since then it has inspired many other transformed based models, for instance RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), and T5 (Raffel et al., 2020) , xlm and XLNet (Yang et al., 2019).

Such models also have the advantage of being able to recognise multiple languages, giving rise to multilingual models and model variants, such as M-BERT, XLM-R (Conneau et al., 2019), mT5 (Xue et al., 2021) and RemBERT (Chung et al., 2020). Nevertheless, the training requires large amounts of training data, which many languages lack, leading to varying performance between different languages. They have also shown to perform worse than monolingual models (Martin et al., 2020; Virtanen et al., 2019). As such Ulčar et al. (Ulčar and Robnik-Šikonja, 2020) made an effort to strike a middle ground between the performance of monolingual and versatility of multilingual models by reducing the number of languages in multilingual model to three; two similar less-resourced languages from the same language family and English. This resulted in two trilingual models FinEst BERT and CroSloEngual BERT al. (Ulčar and Robnik-Šikonja, 2020).

In 2020, a Slovene monolingual RoBERTa-based model SloBERTa (Ulčar and Robnik-Šikonja, 2021) was introduced. It was trained on 5 different corpora, totaling 3.41 billion words. The latest version of the model is SloBERTa 2.0, augmenting the original model by more than doubling the number of training iterations. The authors evaluated its performance on named-entity recognition, part-of-speech tagging, dependency parsing, sentiment analysis and word analogy, but not on question answering.

While the described advancements of natural language processing models already offer us a partial solution for the lack of language-specific training corpora, namely the ability to train the model on a language where large corpora are present (e.g. English), the models still require language-specific fine-tuning, for which a sizable corpora is needed. In our work we present a potential solution, by using the machine-translation methods to translate smaller corpora to Slovene and use it to fine-tune and evaluate the results.

## 3. Dataset description and methodology

**Stanford Question Answering Dataset (SQuAD 2.0)** (Rajpurkar et al., 2018) is a reading comprehension dataset. It is based on a set of articles on Wikipedia which cover a variety of topics, from historical, pharmaceutical, and religious texts to texts about the European Union. Every question in the dataset is a segment of text or span from the corresponding reading passage. It consists of over 100,000 question-answer pairs extracted from over 500 articles.

The reason to use Squad 2.0 over 1.0 is that it consists of twice as much data and contains unanswerable questions.

### 3.1. Machine Translation

To translate the dataset into Slovenian we used the eTranslation webservice (Commission, 2020). Due to the web service being primarily designed to translate webpages and short documents in docx or pdf format, our translation pipeline design was as follows:

1. Convert the corpus in html format.
2. Split html file into smaller chunks. We found that 4 MB chunks work best, as larger chunks were often unable to be translated.
3. Send chunks to the translation service.
4. Use the original corpus file to compose the translated document in the original format.

Since the basic translation yielded quite underwhelming results, we employed two different methods to improve the results. The first was to correct the answers by breaking down both the answer and the context into lemmas and search for the answer sequence of lemmas in context sequence of lemmas. To accomplish this, CLASSLA (CLARIN Knowledge Centre for South Slavic languages) library (Ljubešić and Dobrovoljc, 2019) was used. If a match was found, we replaced the bad answer with the original text, forming the lemma sequence in the context. The second method was to embed the answers in the context before translation.

To evaluate the quality of different translations, we measured how many answers can be directly found in their respective context, as they cannot be used in QA models otherwise. The results can be seen in Table 1. Resulting number of valid questions, compared with the original, are presented in Table 2.

| Basic | LC | CE | LC+CE |
|---|---|---|---|
| 44% | 66% | 93% | 94% |

Table 1: Results for basic translation, lemma correction (LC), and context embedded (CE) translation of SQuAD 2.0 dataset. The percentages represent the number of answers that can be directly found in the respective context.

| Dataset | Subset | AQ | IQ | Total |
|---|---|---|---|---|
| Original | Train | 86,821 | 43,498 | 130,319 |
| | Test | 5,928 | 5,945 | 11,873 |
| Machine Trans. | Train | 81,884 | 43,498 | 125,382 |
| | Test | 5,735 | 5,945 | 11,680 |

Table 2: Number of questions in original SQuAD 2.0 dataset and our machine translated dataset. *AQ* denotes the number of answerable questions, *IQ* the number of impossible questions.

### 3.2. Post-editing of Machine Translation

Due to limited human resources, post-editing was done on small number of automatically translated excerpts that were chosen randomly. The provided excerpts included original paragraphs or contexts, questions and answers, as well as their machine translations, which were to be corrected by a translation student. This was done in two steps: creating a project in the online translation tool Memsource with translation memory in tmx format, generated from machine translations, and revision or post-editing of the segments. Editing was first done on the paragraphs and then on questions and answers, since the answers had to match the text in the paragraph. The editing was minimal, which means that the focus was not on stylistic improvement, but mostly on correcting the grammatical errors, wrong meanings and very unusual syntax, to make the translation comprehensible. As mentioned above, the topics of original texts are diverse and very technical, covering different domains such as religion, history, politics, mathematics and chemistry.

In total, there were 30 manually corrected contexts with accompanying 142 answerable and 143 unanswerable questions. The number of different segment types and of post-editing changes can be seen in Table 3.

| **Segment content** | **S** | **NS** | **CS** | **FS** |
|---|---|---|---|---|
| Context | 30 | 0 | 30 | 100% |
| Answerable question | 142 | 38 | 104 | 73.2% |
| Answer | 435 | 225 | 210 | 48.3% |
| Impossible question | 143 | 43 | 100 | 69.9% |
| **Total number** | 750 | 306 | 444 | 59.2% |

Table 3: Post-editing numerical data. *S* denotes the number of segments, *NS* the number of non-corrected segments, *CS* the number of corrected segments and *FS* the fraction of corrected segments.

### 3.3. Post-editing Analysis

The numbers seen in Table 3 are not fully representative, since some corrections of the mistakes of machine translation are more severe than others and in some segments, there is a much greater number of corrections than in others. For instance, the corrections, including one of a severe semantic mistake, can be seen in this example:

1. Original: The Northern Chinese were ranked higher and Southern Chinese were ranked lower because southern China withstood and fought to the last before caving in.
2. Machine translation: Severna Kitajci so bili uvrščeni višje in južna Kitajci so bili uvrščeni nižje, ker je južna Kitajska zdržala in se borila do zadnjega pred jamarstvom.
3. Post-edited machine translation: Severni Kitajci so bili uvrščeni višje in južni Kitajci so bili uvrščeni nižje, ker se je južna Kitajska pred predajo upirala in se borila do zadnjega.

Answerable and impossible questions have a similar percentage of segments with corrections. This percentage is quite high because machine translation provided incoherent results. In this segments, the changes in post-editing are also more notable, because they affect the overall understanding for potential readers. This can be seen in the following examples:

**Original**

1. Who did Kublai make the ruler of Korea?
2. Who was Al-Banna's assassination a retaliation for the prior assassination of?
3. What plants create most electric power?

**Machine translation**

1. Kdo je Kublai postal vladar Koreje?
2. Kdo je bil Al-Bannin umor maščevanja zaradi predhodnega umora?
3. Katere rastline ustvarjajo največ električne energije?

**Post-edited machine translation**

1. Koga je Kublajkan nastavil za vladarja Koreje?
2. Al-Bannov umor je bil maščevanje za čigav predhodni umor?
3. Katere naprave ustvarjajo največ električne energije?

The segments with answers have the largest number of non-corrected segments because they are shorter. Nevertheless, the percentage of corrected questions is still high if we take into account that the answers represent 58% of all segments. The mistakes in the answers were in the most part already corrected in the contexts. More severe mistakes include semantic mistakes (e.g. plants translated as 'rastline', not 'naprave') and completely wrong answers (e.g. empty segment instead of 'Fermilab' or 'in' instead of '1,388'). Some frequent mistakes also occured in translations of the names of movements, books, projects or other names (e.g. 'Bricks for Varšava' was left untranslated and was changed to 'Zidaki za Varšavo'). There were some punctuation errors, but the most interesting are grammatical mistakes, especially when the wrong grammatical case, gender or number is used. Even if these mistakes were corrected in the context, the answers had to be in the exact same form, so many answers do not sound coherent, which is of course not the case for English, where the conjugation does not change the words as much (e.g. 'Which part of China had people ranked higher in the class system?' — 'Northern' — 'V katerem delu Kitajske so bili ljudje višje v razrednem sistemu?' — 'Severni' (from the example of a sentence in the context mentioned above)). On the other part, some corrected segments were identical even though the source was different due to the use of articles in English language (e.g. 'North Sea' and 'the North Sea' were both translated as 'Severno morje').

It should also be noted that the database SQuAD 2.0 is not entirely reliable. From the batch of randomly sampled 142 test question and answer groups, there were 14 occurrences where at least one of the given answers was not correct (e.g. 'Advanced Steam movement' instead of 'pollution' as an answer to 'Along with fuel sources, what concern has contributed to the development of the Advanced Steam movement?').

# 4. Models

In this section we present each of the five models that were used in the evaluation.

## 4.1. XLM-R

XLM-R (XLM-RoBERTa) (Conneau et al., 2019) is a pre-trained cross-lingual language model based on xlm (Lample and Conneau, 2019). The 'RoBERTa' part of the name comes from its training routine that is the same as the monolingual RoBERTa model, specifically, that the sole training objective is the MLM (masked language mode). There is no next sentence prediction (as in BERT) or Sentence Order Prediction (as in ALBERT). XLM-R shows the possibility of training one model for many languages while not sacrificing per-language performance. It is trained on 2.5 TB of CommonCrawl data in 100 languages.

## 4.2. M-BERT

M-BERT (Multilingual Bert) (Devlin et al., 2018) is a pre-trained cross-lingual language model as it's name suggest. It is based on BERT (Devlin et al., 2018). The pre-trained model is trained on 104 languages with large amount of data from Wikipedia, using a masked language modeling (MLM) objective. On Hugging Face, there is only a base model with 12 hidden transformer layers available, large model with 24 hidden transformer layers was not uploaded and we were not able to test it.

## 4.3. RemBERT

RemBERT (Chung et al., 2020) is a model, pre-trained on 110 languages, using a masked language modeling (MLM) objective. It's difference with mBERT is that the input and output embeddings are not tied. Instead, RemBERT uses small input embeddings and larger output embeddings. This makes the model more efficient since the output embeddings are discarded during fine-tuning.

## 4.4. SloBERTa

SloBERTa (Ulčar and Robnik-Šikonja, 2021) is a Slovene monolingual large pre-trained masked language model. It is closely related to French Camembert model, which is similar to base RoBERTa model, but uses a different tokenization model. Since the model requires a large dataset for training, it was trained on 5 combined datasets. It outperformed existing Slovene models.

## 4.5. CroSloEngual BERT

It is a trilingual model based on BERT and trained for Slovene, Croatian and English language. It was trained with 5.9 billion tokens from these languages. For those languages it performs better than multilingual BERT, which is expected, since studies showed that monolingual models perform better than large multilingual models (Virtanen et al., 2019).

# 5. Results

This section is divided into two parts. First we evaluate automatic machine translations and then we evaluate performance of choosen QA models (XLM-R-large, M-Bert-base, CroSloEngual BERT, RemBERT, SloBERTa 2.0). All tests were performed on i5 10400f system with RTX 3070 GPU 8 GB VRAM. For larger models we used RTX 3060 12 GB.

To compare the performance between the English, machine translated Slovene and human translated Slovene versions of the SQuAD 2.0 dataset, we used 5 different question answering models: mBERT, XLM-R, RemBERT, SloBERTa 2.0, CroSloEngual BERT. The evaluation was done in three steps:

1. Performance evaluation of different models and fine-tuning configuration on the English dataset, as a benchmark for the evaluation of the Slovene results.
2. Performance evaluation of different models and fine-tuning configuration on the Slovene dataset, translated using computer only, to evaluate the quality of machine translation.
3. Performance evaluation of different models and fine-tuning configuration on the Slovene subset which was translated by a human, and same subset both in English and translated using computer, to evaluate the benefits of human translation.

Before the evaluation, we removed all punctuation, leading and trailing white spaces and articles from both ground truth and prediction. Both of them were also set in the lower case. Parameters used for fine-tuning are presented in Table 4.

Metrics used for the evaluation match the official ones for SQuAD2.0 evaluation and were as follows:

- **Exact** - The fraction of predictions matched at least of one the correct answers exactly.
- **F1** - The average overlap between prediction and ground truth, defined as an average of F1 scores for individual questions. F1 score of an individual question is computed as a harmonic mean of the precision and recall, where precision was defined as $\frac{T_M}{T_P}$, and recall as $\frac{T_M}{T_{GT}}$, where $T_M$ represents the matching tokens between prediction and ground truth, $T_P$ number of tokens in prediction and $T_{GT}$ number of tokens in ground truth. A token is defined as a word, separated by a white space.

The results of the non-translated SQuAD 2.0 and machine translated dataset can be seen in Table 5. The results of the human translated subset and its English and computer translated counterparts can be seen in Table 6. Additionally, we provide some examples of correct predictions with wrong answers in Table 7 and some of correct answers with wrong predictions in Table 8.

| Model Name | B | MS | LR | E |
|---|---|---|---|---|
| XLM-R-large | 4 | 256 | 1e-5 | 3 |
| M-BERT-base | 8 | 320 | 3e-5 | 3 |
| CroSloEngual BERT | 4 | 256 | 1e-5 | 3 |
| RemBERT | 4 | 256 | 1e-5 | 3 |
| SloBERTa 2.0 | 16 | 320 | 3e-5 | 3 |

Table 4: Parameters used to fine-tune the evaluated models. *B* denotes the number of batches used during fine-tuning, *MS* the maximum sequence length, *LR* the learning rate and *E* the number of epochs.

| Model name | Fine-Tuning Language | Original | | Machine Translation | |
|---|---|---|---|---|---|
| | | Exact | F1 | Exact | F1 |
| xlmR-large | Eng | 81.8% | 84.9% | 64.3% | 72.3% |
| xlmR-large | Slo | 75.0% | 79.2% | 65.3% | 72.4% |
| xlmR-large | Eng & Slo | 74.4% | 78.5% | 65.9% | 73.4% |
| M-BERT-base | Eng | 75.6% | 78.9% | 55.4% | 61.3% |
| M-BERT-base | Slo | 62.4% | 67.2% | 60.4% | 67.0% |
| M-BERT-base | Eng & Slo | 70.7% | 75.0% | 60.5% | 67.3% |
| CroSloEngual BERT | Eng | 72.8% | 76.3% | 56.3% | 63.6% |
| CroSloEngual BERT | Slo | 63.6% | 68.2% | 58.4% | 65.4% |
| CroSloEngual BERT | Eng & Slo | 68.8% | 73.0% | 58.1% | 65.7% |
| RemBERT | Eng | **84.5%** | **87.5%** | **67.1%** | **73.8%** |
| SloBERTa 2.0 | Slo | 60.6% | 64.7% | **66.7%** | **73.9%** |

Table 5: Comparison of the results of various models and their fine-tuning configurations on the English SQuAD 2.0 evaluation dataset and Slovene machine translated SQuAD 2.0 evaluation dataset. The English dataset only contains the questions preset in its Slovene counterpart. Specific parameters used in fine-tuning are presented in Table 4.

| Model name | Fine-Tuning Language | Original | | Machine Translation | | Human Translation | |
|---|---|---|---|---|---|---|---|
| | | Exact | F1 | Exact | F1 | Exact | F1 |
| xlmR-large | Eng | 80.0% | 82.9% | 61.1% | 68.5% | 71.6% | 75.9% |
| xlmR-large | Slo | 69.1% | 72.9% | 61.4% | 69.1% | 69.8% | 74.8% |
| xlmR-large | Eng & Slo | 68.8% | 73.4% | 64.6% | 72.4% | 70.5% | 75.7% |
| M-BERT-base | Eng | 71.9% | 74.9% | 52.6% | 57.7% | 57.5% | 60.3% |
| M-BERT-base | Slo | 56.1% | 60.4% | 58.6% | 64.5% | 60.4% | 66.2% |
| M-BERT-base | Eng & Slo | 64.9% | 68.8% | 55.8% | 61.2% | 63.5% | 68.6% |
| CroSloEngual BERT | Eng | 73.3% | 75.5% | 53.0% | 60.8% | 62.1% | 65.7% |
| CroSloEngual BERT | Slo | 59.6% | 63.1% | 51.6% | 58.8% | 60.7% | 66.0% |
| CroSloEngual BERT | Eng & Slo | 68.1% | 70.6% | 58.9% | 66.3% | 64.6% | 71.0% |
| RemBERT | Eng | **84.9%** | **87.2%** | 64.2% | 71.4% | 71.9% | 76.9% |
| SloBERTa 2.0 | Slo | 59.3% | 65.0% | **64.9%** | **72.2%** | **72.6%** | **78.0%** |

Table 6: Comparison of the results of various models and their fine-tuning configurations on the Human Translated subset of SQuAD 2.0, and the subsets containing same question from original English dataset and the machine translated dataset. Specific parameters used in fine-tuning are presented in Table 4.

| # | Dataset | Question | Answer | Prediction |
|---|---|---|---|---|
| 1 | ENG | How many of Warsaw's inhabitants spoke Polish in 1933? | 833,500 | 833,500 |
| | MT | Koliko prebivalcev Varšave je leta 1933 govorilo poljsko? | prebivalcev | 833.500 |
| | HT | Koliko prebivalcev Varšave je leta 1933 govorilo poljski jezik? | 833.500 | 833.500 |
| 2 | ENG | Who recorded "Walking in Fresno?" | Bob Gallion | Bob Gallion |
| | MT | Kdo je posnel „Walking in Fresno?“ | je Bob | Bob Gallion |
| | HT | Kdo je posnel ≫Walking in Fresno≪? | Bob Gallion | Bob Gallion |

Table 7: Examples of correct predictions with wrong answers. *ENG* denotes the English dataset, *MT* one translated by a computer and *HT* one translated by a human.

| # | Dataset | Question | Answer | Prediction |
|---|---|---|---|---|
| 1 | ENG | Where did Korea border Kublai's territory? | northeast | northeast |
| | MT | Kje je Koreja mejila na Kublajevo ozemlje? | severovzhodno | zahodno |
| | HT | Kje je Koreja mejila na Kublajkanovo ozemlje? | severovzhodno | severovzhodno |
| 2 | ENG | How many miles, once completed, will the the Lewis S. Eaton trail cover? | 22 | 22 |
| | MT | Koliko kilometrov, ko bo končano, bo pokrivalo Lewis S. Eaton? | 22 | 35 |
| | HT | Koliko kilometrov bo dolga pot Lewisa S. Eatona, ko bo končana? | 22 | 35 |

Table 8: Examples of correct answers with wrong predictions. *ENG* denotes the English dataset, *MT* one translated by a computer and *HT* one translated by a human.

# 6. Discussion

## 6.1. Quantitative Analysis

From the results in Table 5, we can see that RemBERT and SloBERTa 2.0 gave the best results on the dataset translated by a computer. While the result for SloBERTa was expected, as monolingual models tend to perform better than multilingual ones, RemBERT managed to outperform its multilingual competitors while only being fine-tuned on the English dataset. We would attribute this simply to the better design of the model. Although both models had a very similar performance, we would like to point out that RemBERT model is a much larger model and was pre-trained on a significantly larger dataset. Similar results were also observed when comparing the results on the smaller subset of questions that were translated by a human, as seen in Table 6.

In Table 6 we can see models consistently performing better on the human translated data, suggesting that the machine translation provided by eTranslation webservice comes short of providing adequate set for proper evaluation in the Slovene language. We can also see that while the models fine-tuned using machine translated dataset do perform better when evaluated on the machine translated data, this does not hold true for evaluations on human translated data.

We have also observed that fine-tuning the model on the English dataset first, and then on the Slovene, yields better results on the smaller models, M-BERT-base and CroSloEngular BERT, as compared to fine-tuning on either language.

## 6.2. Qualitative Analysis

While there are many correct predictions of the answers in the machine translated dataset, it is clear that a great number of predictions still does not answer the question correctly. This is because the machine translation of the sentences in the context is not grammatically and stylistically correct, does not convey the right meaning and thus the model has more problems finding the answer. The correct predictions are mostly the ones where the answer to the question is short and the words are not conjugated, i.e. numbers and names, even though there are some exceptions. The same is true for human post-edited translation, but improvement of some answers is already visible from only a few representative examples in Table 7 and Table 8.

# 7. Conclusion

In this work we present a machine translated SQuAD 2.0 dataset and evaluate it on the following question answering (QA) models: XLM-R-large, M-BERT-base, RemBERT, CroSloEngual BERT and SloBERTa 2.0. Additionally, we also perform human post-editing on a subset of SQuAD 2.0 translations in order to better ascertain the quality of machine translations. The results show that using machine translated data for evaluation led to notably worse results as compared to the one translated by a human. Moreover, we noticed that while multilingual models fine-tuned using machine translated data performed better than ones fine-tuned on English data when given a task of answering the machine translated question, the situation was in most cases reversed when given a task of answering human translated questions. This leads us to conclude that machine translation, at least one available on via eTranslation (Commission, 2020) service, is not particularly suitable for training multilingual models. Of all the models, SloBERTa 2.0 produced the best results on both machine and human translated data, while the RemBERT gave comparable results even when only fine-tuned on the English dataset.

The testing procedure could be easily improved by employing stronger hardware. RemBERT could for example be fine-tuned on the Slovene dataset, which would allow for its better evaluation. Additionally, we were unable to ascertain the optimal parameters for fine-tuning as performing multiple fine-tunings for each language would be unfeasible. Some restrictions of the project are limited time for post-editing and only one translator who is not an expert in the topics of various technical texts, and the method of minimal editing that can result in mediocre translation. The experiment could be expanded by including a larger subset of human translated or revised data, more datasets, such as Natural Questions (Kwiatkowski et al., 2019), and different machine translation services, such as DeepL.

# 8. Acknowledgments

We would like to to thank our mentors, Slavko Žitnik and Špela Vintar, for providing us with directions, feedback and advice.

# 9. References

Ines Čeh and Milan Ojsteršek. 2009. Developing a question answering system for the Slovene language. *WSEAS Transaction on Information science and applications*, (9).

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *CoRR*, abs/2010.12821.

European Commission. 2020. CEF Digital eTranslation. `https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation`.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. *Unsupervised cross-lingual representation learning at scale. arXiv:1911.02116.*

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In: *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In: *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In: *Proceedings of the 58th Annual Meeting of the Association for Computa- tional Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad.

Matej Ulčar and Marko Robnik-Šikonja. 2020. Finest BERT and CroSloEngual BERT. In: *International Conference on Text, Speech, and Dialogue*, pages 104–111. Springer.

Matej Ulčar and Marko Robnik-Šikonja. 2021. SloBERTa: Slovene monolingual large pretrained masked language model.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Fnnish. *arXiv:1912.07076*.

William A Woods and WOODS WA. 1977. Lunar rocks in natural english: Explorations in natural language question answering.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pretrained text-to-text transformer. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.