Adapting an English Corpus and a Question Answering System for Slovene

Uroš ŠMAJDEK
Faculty of Computer and Information Science, University of Ljubljana

Matjaž ZUPANIČ
Faculty of Computer and Information Science, University of Ljubljana

Maj ZIRKELBACH
Faculty of Computer and Information Science, University of Ljubljana

Meta JAZBINŠEK
Faculty of Arts, University of Ljubljana

Developing effective question answering (QA) models for less-resourced languages like Slovene is challenging due to the lack of proper training data. Modern machine translation tools can address this issue, but this presents another challenge: the given answers must be found in their exact form within the given context since the model is trained to locate answers and not generate them. To address this challenge, we propose a method that embeds the answers within the context before translation and evaluate its effectiveness on the SQuAD 2.0 dataset translated using both eTranslation and Google Cloud translator. The results show that by employing our method we can reduce the rate at which answers were not found in the context from 56% to 7%. We then assess the translated datasets using various transformer-based QA models, examining the differences between the datasets and model configurations. To ensure that our models produce realistic results, we test them on a small

Šmajdek, U., Zupanič, M., Zirkelbach, M., Jazbinšek, M. Adapting an English Corpus and a Question Answering System for Slovene. Slovenščina 2.0, 11(1): 247–274.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2023.1.247-274

subset of the original data that was human-translated. The results indicate that the primary advantages of using machine-translated data lie in refining smaller multilingual and monolingual models. For instance, the multilingual CroSloEngual BERT model fine-tuned and tested on Slovene data achieved nearly equivalent performance to one fine-tuned and tested on English data, with 70.2% and 73.3% questions answered, respectively. While larger models, such as RemBERT, achieved comparable results, correctly answering questions in 77.9% of cases when fine-tuned and tested on Slovene compared to 81.1% on English, fine-tuning with English and testing with Slovene data also yielded similar performance.

Keywords: question answering, machine translation, multilingual models

1 Introduction

One of the goals of artificial intelligence is to build intelligent systems that can interact with humans and help them. One such task is reading the web and then answering complex questions about any topic with regard to the given context. These question answering 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.

Question answering (QA) is one of the disciplines in the broader field of natural language processing (NLP), which involves the automatic answering of questions posed in natural language. Thus, the goal of QA is the development of automated systems that can understand and respond to human questions in a way that is similar to how humans answer questions. The QA task is typically formulated as follows: given a question and a context, the system must identify the answer to the question within the given context.

Recently, pre-trained contextual embedding (PCE) models like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) have attracted plenty of attention due to their good performance in a wide range of NLP tasks, including QA. Compared to earlier

information retrieval and knowledge-based systems, modern QA systems are significantly less domain-dependent, as they do not require a specifically tailored database to function effectively. This has thus led to the development of multilingual question answering systems, where the same system can serve a multitude of languages.

However, multilingual QA tasks typically assume that answers exist in the same language as the question, and require a smaller corpus to fine-tune it for a given language and a broader domain (e.g. Wikipedia articles). 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 not feasible, and therefore we often employ machine translations instead. However, machine translation has trouble interpreting the nuances of specific languages, such as culturally specific vocabulary (e.g. translating bird sanctuary as "ptičje zatočišče", where the correct translation is "ptičji rezervat"), the use of articles, proper nouns, abbreviations, and implicit relationships between words (Koehn and Knowles, 2017; Arnejšek and Unk, 2020). This is especially problematic in question answering, where the answer has to be found in its exact form within the context to be usable for training such a model.

The objective of our work is thus to reduce the impact of errors in the construction of a machine-translated (MT) dataset that can be used to both fine-tune and test a question answering (QA) model. Specifically, we focus on the translation of the popular SQuAD 2.0 (Rajpurkar et al., 2018) QA dataset. Moreover, we benchmark the accuracy of QA models fine-tuned using the proposed MT dataset by assessing them on a human-translated (HT) subset of the original data.

The main contributions of our work are:

- a pipeline for translation of an English QA dataset;
- performance comparison of the various monolingual and multilingual QA models fine-tuned on the original dataset and the English-to-Slovene MT datasets;
- comparison of the eTranslation and Google Cloud Translation services in terms of raw translation and QA performance using the data translated from English to Slovene; and

- evaluation of the QA performance of the resulting QA models on the HT subset.

This paper is a follow-up to our submission to JDTH 2022 (Zupanič et al., 2022). To improve upon the presented concept, we expanded our evaluation to include the corpus translated by the state-of-the-art Google Cloud Translation (Google CT) service to assess the impact of the quality of translations. In addition, to ensure the testing set is not influenced by the machine translation, we replaced the post-edited machine translation samples with a fully human-translated testing set. Lastly, we also experimented with additional model parameters during evaluation and improved the presentation of our method.

In Section 2 we present the related work. In Section 3 we present our dataset, the process of translation, 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 & WOODS, 1977), date back to the 1960s and the 1970s. They were characterized 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 systems 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 in the Slovene language was presented by Čeh and Ojsteršek (2009), where the authors used classification and answer retrieval in parallel. The system retrieved data from its own database, consisting of MS Excel files, local databases, and integrated web services. For question classification, they used Support Vector Machines. The problem was that the system was very limited in question answering, able to answer only specific predefined classes of questions.

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., 2019), a transformer model introduced in 2018. Since then it has inspired many other transformer-based models, such as RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), and T5 (Raffel et al., 2020), XLM (Lample and Conneau, 2019) and XLNet (Yang et al., 2019).

Such models also have the advantage of being able to recognize multiple languages, giving rise to multilingual models and model variants, such as Multilingual BERT, XLM-RoBERTa (Conneau et al., 2020), mT5 (Xue et al., 2021) and RemBERT (Chung et al., 2021). Nevertheless, the training requires large amounts of training data, which many languages lack, leading to varying performance between different languages. They have also been shown to perform worse than monolingual models (Martin et al., 2020; Virtanen et al., 2019). Ulčar and Robnik-Šikonja (2020) thus made an effort to strike a middle ground between the performance of monolingual models and the versatility of multilingual ones by reducing the number of languages in the 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.

In 2020, a Slovene monolingual RoBERTa-based model called SloBERTa (Ulčar et al., 2021) was introduced. It was trained on five different corpora, totalling 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 advances 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 corpus is needed. In our work, we present a potential solution to this problem, by using

machine-translation methods to translate smaller corpora to Slovene, and then fine-tune and evaluate the results.

3 Dataset description and methodology

In this section, we describe the dataset used in our study and the methodology employed to create machine- and human-translated datasets for fine-tuning and testing the question answering models.

3.1 Stanford Question Answering Dataset (SQuAD 2.0)

SQuAD 2.0 (Rajpurkar et al., 2018) is a reading comprehension dataset. It is based on a set of articles on Wikipedia that cover a variety of topics, from historical, pharmaceutical, and religious texts to texts about the European Union. Every question in the dataset is associated with 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. Unlike SQuAD 1.0, the dataset contains roughly twice as much data, and also includes unanswerable questions, which are designed to look similar to answerable ones, but lack an answer within the given text. Thus, for a system to perform well on SQuAD2.0, it must not only answer questions, when possible but also determine when no

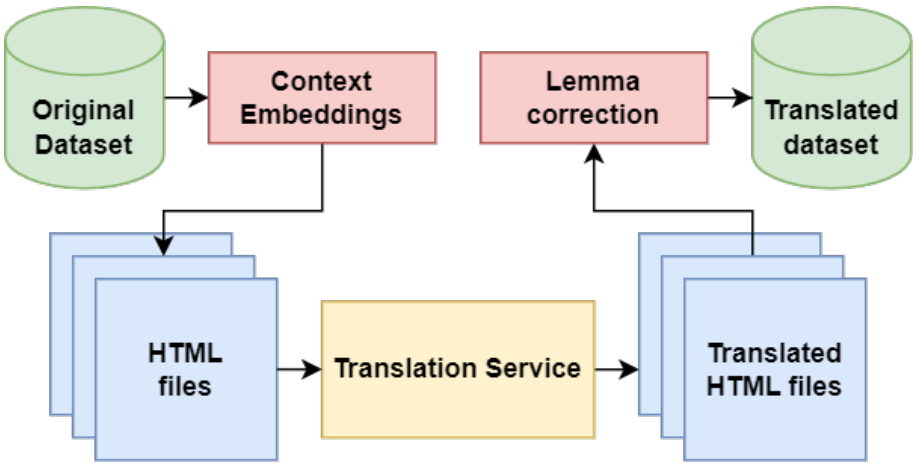

Figure 1: An overview of the machine translation pipeline.

answer is supported by the paragraph and abstain from answering. An example of a question-answer pair from the dataset is:

- **Question:** What is the name of the state referred to by historians during the Middle Ages as the Eastern Roman Empire?
- **Context:** During the Middle Ages, the Eastern Roman Empire survived, though modern historians refer to this state as the Byzantine Empire...
- **Answer:** the Byzantine Empire

3.2 Machine translation

In this subsection, we will describe the proposed machine-translation pipeline, a brief overview of which can be found in Figure 1. To translate the dataset into Slovenian we used two translation web services: eTranslation (European Commission, 2020) and Google CT. Due to the web services 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 into HTML format. We wrapped context-question-answer coupling in separate HTML tags and placed them within a hierarchy resembling that in the original dataset. HTML tag attributes were used to preserve the unique question and answer identifiers for later evaluation. An example of the resulting structure can be seen in Appendix B.
2. Split the 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.

The requirement for a context-question-answer coupling to be used to train a question answering model is that the answer has to be found in its exact form within the context. For example, if the answer entity is *Bizantinsko cesarstvo*, but the relevant part of the context was translated as *...v Bizantinskem cesarstvu...*, such a question-answer pair is unusable. This is due to the model's task being to find the index

of an answer to the question within the context. To improve upon basic translation, we employed two different methods.

The first was to correct the answers by breaking down both the answer and the context into lemmas and searching for the lemmatized form of the answer in the lemmatized form of the context. To accomplish this, the CLASSLA (CLARIN Knowledge Centre for South Slavic languages) language processing pipeline (Ljubešić and Dobrovoljc, 2019) was used. If a match was found, we replaced the bad answer with the matching original text in the context.

The second method was to embed the answers in the context before translation. This was done by replacing the answer entry of the untranslated document with a copy of the context entry which had the answer marked by a common HTML tag and a unique attribute to avoid mistaking it for a preexisting tag within the context. This allows the translation to also take into account the context surrounding the answer, greatly increasing the chance such an answer will be found in the original context. As the locations of answers within contexts are given by the dataset, finding the correct context entry is a trivial operation. For example, the untranslated answer entry '*the Byzantine Empire*' was replaced with the following: '*During the Middle Ages, the Eastern Roman Empire survived, though modern historians refer to this state as the Byzantine Empire...*'.

3.3 Human translation

When we added the Google CT service, we replaced the post-editing with the completely human translation of the excerpts so that the end comparison would be more objective and of better quality. The translation was done on a small number of automatically translated excerpts chosen randomly due to limited human resources.

The provided excerpts included original paragraphs or contexts, questions, and answers. Firstly, we translated the paragraphs and then the questions and answers since the answers had to match the text in the paragraph. As mentioned above in the description of the dataset, the topics of the original texts were very diverse and technical, covering different domains such as religion, history, politics, mathematics, and chemistry.

In total, there were 30 translated contexts with accompanying answerable and unanswerable questions, as well as impossible questions. The exact number of different segment types can be seen in Table 1.

Table 1: *Statistics for manually translated data subset*

Segment content	Number of instances
Context	30
Answerable question	142
Answer	435
Impossible question	143
Total number	**750**

4 Models

In this section, we present each of the five models that were used in the evaluation (Table 2). Since those transformer models are usable for a various number of natural language tasks we used Hugging Face's question answering pipeline to infer with question answering models.[1] The following models were selected as they are diverse in terms of properties and are publicly available, well documented, and have shown promising performance figures in the past.

Table 2: *Used models with the respective properties*

Model Name	Trained languages	No. of training tokens	No. of hidden layers
SloBERTa	1	3.47 billion	12
CroSloEngual BERT	3	5.9 billion	12
Multilingual BERT	104	No data	12
XLM-RoBERTa	100	6.3 trillion	24
RemBERT	110	1.8 trillion	32

4.1 SloBERTa

SloBERTa (Ulčar & Robnik-Šikonja, 2021) is a Slovene monolingual large pre-trained masked language model. It is closely related to the French CamemBERT model, which is similar to the base RoBERTa with

1 https://huggingface.co/docs/transformers/tasks/question_answering

12 hidden transformer layers but uses a different tokenization model. Since the model requires a large dataset for training, it was trained on five combined datasets with 3.47 billion tokens. It outperformed existing Slovene models.

4.2 CroSloEngual BERT

CroSloEngual BERT (Virtanen et al., 2019) is a trilingual model based on a BERT base with 12 hidden transformer layers and trained for Slovene, Croatian, and English, using 5.9 billion tokens from these languages. For those languages it performs better than multilingual BERT, which is expected, since studies show that monolingual models perform better than large multilingual ones (Virtanen et al., 2019).

4.3 Multilingual BERT

Multilingual BERT (M-BERT) (Devlin et al., 2019) is a version of BERT that has been trained on data from Wikipedia in 104 languages to satisfy the demand for multilingualism. It can perform tasks such as question answering, language classification, and many more for a wide range of languages.

The model was released in large form with 24 hidden transformer layers and in base form with 12 hidden transformer layers. We used the latter one in the current work.

4.4 XLM-RoBERTa

XLM-RoBERTa (XLM-R) (Conneau et al., 2020) is a pre-trained cross-lingual language model developed by Facebook AI. It is trained on 2.5 TB of CommonCrawl data, with a total of 6.3 trillion tokens in 100 languages, and based on RoBERTa (Robustly Optimized BERT Pretraining Approach) (Lample and Conneau, 2019). XLM-R, like M-BERT, uses a similar pretraining objective. However, XLM-R has a larger model size, shared vocabulary, and is trained using more training data from the web. XLM-R large, which we used in our work, has 24 hidden layers.

4.5 RemBERT

RemBERT (Chung et al., 2021) is a pre-trained multilingual model using a masked language modelling (MLM) objective. This model is pre-trained on 1.8 trillion tokens in 110 languages and is similar to mBERT. However, it differs in that its input and output embeddings are not tied. Instead, it uses small input embeddings and larger output embeddings, which makes the model more efficient because the output embeddings are discarded during fine-tuning. RemBERT, which has 32 hidden transformer layers, is the largest model we tested.

5 Evaluation and results

In this section, we present the evaluation results of our machine translation methods and the performance of the question answering models fine-tuned on the translated datasets.

5.1 Machine translation

To evaluate the quality of different translation methods, we measured how many answers can still be found within their respective context in their exact form. The results for the eTranslation service can be seen in Table 3. The resulting number of valid questions for both translation services, compared with the original dataset, are presented in Table 4.

Table 3: *Results for basic translation, lemma correction (LC), and context embedded (CE) translation of SQuAD 2.0 dataset by eTranslation*

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

Note. The percentages represent the number of answers that can be found within their respective context in their exact form.

Table 4: *Number of questions in the original SQuAD 2.0 dataset and our machine-translated datasets*

Dataset	Subset	AQ	AQ [%]	IQ	Total
Original	Train	86,821	66.6%	43,498	130,319
	Test	**5,928**	49.9%	5,945	11,873
eTranslation	Train	81,884	65.3%	43,498	125,382
	Test	**5,735**	49.1%	5,945	11,680
Google CT	Train	84,048	65.9%	43,498	127,546
	Test	**5,821**	49.5%	5,945	11,766

Note. AQ denotes the number of answerable questions and IQ the number of impossible questions.

5.2 Question answering

In order to evaluate the question answering performance of MT datasets obtained by eTranslation and Google CT, we first used both of them and the original English dataset, to fine-tune the following question answering models: M-BERT, XLM-R, RemBERT, SloBERTa 2.0, CroSloEngual BERT. This yielded 14 different fine-tuned model configurations, as showcased in the first two columns of Table 6. The fine-tuned models were then evaluated in two stages described later in the section. All tests were performed on a system with an Intel Xeon E5-2687W v4 @ 3.00GHz CPU and RTX 3090 24GB GPU. 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 lowercase. The parameters used for fine-tuning are presented in Table 5.

The metrics used for the evaluation matched the official ones for the SQuAD 2.0 evaluation and were as follows:

- **Exact** - The fraction of predictions that 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. The F1 score of an individual question is computed as a harmonic mean of the precision and recall, where precision was defined as T_M/T_P, and recall as T_M/T_{GP}, where T_M represents the maching tokens between

prediction and ground trouth, T_P number of tokens in prediction and T_{GP} number of tokens in the ground truth. A token is defined as a word, separated by white space.

In the first step of the evaluation, each of the models was tested using the HT subset of the original English testing set, and its untranslated counterpart. Additionally, we also repeated each of the tests on MT testing subsets, in order to assess their viability to be used as testing sets. The F1 scores can be seen in Table 6, whereas a full result overview is presented in Appendix C, Table 10.

In the second step of the evaluation, we repeated the tests with full original and MT testing sets to account for the potential discrepancies between the difficulty of the original dataset and the subset in the first set of experiments. The results of this step can be seen in Table 7 and Appendix C, Table 11. By comparing this set of results with the ones obtained in the first step we can see that the full set gives slightly better results, implying that the questions chosen in the subset were more difficult on average.

Table 5: *Parameters used to fine-tune the evaluated models*

Model Name	B	MS	LR	E
XLM-R Large	4	265	1e-5	3
M-BERT	8	256	1e-5	3
CroSloEngual BERT	8	320	1e-5	3
RemBERT	4	256	1e-5	3
SloBERTa 2.0	8	320	1e-5	3

Note. 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.

6 Discussion

6.1 Quantitative analysis

By comparing the results of matching entries in Tables 6 and 7, we can observe that the results are consistently better when using the entire dataset instead of the randomly chosen subset for testing. The

differences are relatively minor though and both tables still show the same trends, which we would interpret as a positive indicator of the results in Table 6 being a good representation of the behaviour of the entire dataset.

6.1.1 Manual versus machine translation

By comparing the results of tests using HT data of a model fine-tuned on MT data, with the same model fine-tuned on original data (Table 6) we can see that the impact of fine-tuning with MT datasets varies depending on the size and the inherent performance of the model. The largest performance gain from fine-tuning a multilingual model on MT data as opposed to original data can be observed with M-BERT (F1 score of 74.0% as opposed to 58.2%) and CroSloEngual BERT (F1 score of 73.6% as opposed to 65.1%). However, for the latter, this is only true when using the set translated by Google CT. The impact is less noticeable for the larger RemBERT (F1 score of 81.5% as opposed to 79.5%) model, while worse for the XLM-R Large (F1 score of 81.4% as opposed to 81.6%) model. We reason this to be the result of the inherent ability of those two models to perform well when trained and tested on two different languages. This is clearly visible in our case as these two models, unlike smaller ones, retained much of their ability to answer English questions even when fine-tuned on MT data. It is harder to evaluate the impact of using the MT fine-tuning set on SloBERTa 2.0 since we cannot benchmark it against the same model fine-tuned on the original data, but comparing its results to other similarly sized models – M-BERT and CroSloEngual BERT – we can see that it outperforms both of them, which we would consider as a positive indicator for the viability of using MT data to fine-tune the model. Taking all of this into account we would conclude that using MT data to fine-tune question answering models is a superior option to using original English data, especially if one is constricted to using small models. Additionally, it has the benefit of enabling the use of Slovene monolingual models, which tend to outperform multilingual models of the same size.

6.1.2 Comparison of translation services

Comparing the results of the models fine-tuned with data translated by eTranslation and Google CT and tested on HT data in Table 6, we can observe that in most cases Google CT outperforms eTranslation. The exception to this is M-BERT, where eTranslation slightly outperformed its counterpart, but the difference is not significant. The magnitude of the differences varies from almost insignificant with M-BERT and XLM--R Large, to noticeable with CroSloEngual BERT and RemBERT. We suspect this is due to the inherent differences in the model structure and the data they were pre-trained on.

By observing the results of different models and their variations when tested on the data translated by MT as opposed to the results obtained by testing on HT data (Table 6), we can see that the results vary significantly. We suspect that this is due to the various grammatical errors present in the MT data, as shown in Section 6.2, which are not present in the data that was used to pre-train the models, and thus the models have a harder time recognizing the structure of the context. This is further reinforced by the fact that models consistently yield better results when tested using data translated by Google CT as opposed to eTranslation (Tables 6 and 7) since the former contains fewer grammatical and semantic mistakes (Section 6.2). Additionally, we can also observe a bias where models fine-tuned with one translation service perform better when tested on the same translation service as compared to the model fine-tuned with the other translation service tested on that testing set. All this points us toward concluding that while the MT datasets are a viable solution for fine-tuning they are less suitable for testing, especially if the resulting translation is of lower quality, such as when using the eTranslation service.

Table 6: *Comparison of the F1 scores of various models and their fine-tuning configurations on the human-translated subset of SQuAD 2.0 (N=285), and the subsets containing the same question from the original English dataset and the two machine-translated datasets*

Model Name	Fine-Tuning Dataset	Original [%]	eTranslation [%]	Google CT [%]	Human Transl. [%]
M-BERT	Original	78.4	48.2	55.9	58.2
	eTranslation	62.6	64.5	73.4	74.0
	Google CT	65.1	64.8	71.4	73.6
CroSloEngual BERT	Original	75.5	60.8	63.4	65.1
	eTranslation	63.1	58.8	64.5	63.6
	Google CT	58.8	66.5	66.6	73.6
SloBERTa 2.0	eTranslation	65.0	72.2	76.1	74.9
	Google CT	65.2	68.0	72.9	78.3
XLM-R Large	Original	85.5	69.1	75.8	**81.6**
	eTranslation	82.6	**73.1**	76.9	81.1
	Google CT	82.3	70.9	**77.4**	81.4
RemBERT	Original	**87.2**	71.4	74.3	79.5
	eTranslation	84.1	72.9	76.6	78.6
	Google CT	84.8	71.6	76.0	**81.5**

Note. Specific parameters used in fine-tuning are presented in Table 5.

Table 7: *Comparison of F1 scores of various models and their fine-tuning configurations on the English SQuAD 2.0 evaluation dataset and the two Slovene machine-translated SQuAD 2.0 evaluation datasets (N=11.680)*

Model Name	Fine-Tuning Dataset	Original [%]	eTranslation [%]	Google CT [%]
M-BERT	Original	78.9	59.2	61.9
	eTranslation	68.2	68.3	70.7
	Google CT	68.9	67.9	71.3
CroSloEngual BERT	Original	76.3	63.5	66.8
	eTranslation	68.2	65.5	68.3
	Google CT	65.7	66.5	70.0
SloBERTa 2.0	eTranslation	64.7	73.7	76.8
	Google CT	66.9	72.8	77.0
XLM-R Large	Original	86.3	74.8	78.5
	eTranslation	83.0	**75.6**	78.3
	Google CT	84.4	**75.5**	**80.1**
RemBERT	Original	**87.5**	71.4	74.3
	eTranslation	83.9	72.9	76.6
	Google CT	84.5	71.6	76.0

Note. The English dataset only contains the questions pre-set in its Slovene counterpart. Specific parameters used in fine-tuning are presented in Table 5.

6.2 Qualitative analysis of translations

A comparison of the differences and the types of mistakes in the two machine translations and the human translation was made. The representativeness of these differences cannot be determined, but by looking at more examples some general mistakes of machine translations can be noted. It should also be considered that some of the mistakes of machine translation are more severe than others, and that in some segments there is a much greater number of them than in others, so the mistakes could not be counted exactly.

Firstly, the segments with contexts are very long and this normally led to more grammar, syntactic and stylistic mistakes in machine translations. The eTranslation MT yielded the worst results, as can be seen in Appendix A, example 1, as there was a wrong gender agreement and a big semantical mistake ('caving in' translated as 'jamarstvo'), which did not occur with Google CT. This was expected to a certain degree, as Google CT uses state-of-the-art translation methods, while eTranslation does not. Additionally, eTranslation is designed to perform best when working with texts on EU-related matters (European Commission, 2020) while our dataset is comprised of technical texts which cover a wide variety of topics.

There was also a great dissimilarity between the translations of answerable and impossible questions, because machine translation provided incoherent results. The changes are more notable because they affect the overall understanding of potential readers. These segments are shorter, but in both MT examples the word 'plants' was translated literally, so we can see in the example in the Appendix A that the HT translation is still the best one. Nevertheless, there was a larger number of instances where Google CT performed better than eTranslation at the grammatical and syntactical levels.

The segments with answers were the most similar ones, most probably because they are shorter. The contextual mistakes in the answers were for 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 occurred in translations of the names of movements, books, projects, or other names (e.g. 'Bricks for Varšava' was left untranslated by

eTranslation MT, Google CT did translate it to 'Opeke za Varšavo' and was changed in HT to 'Zidaki za Varšavo'). There were some punctuation errors, but the most interesting are grammatical mistakes of both MT services, especially when the wrong grammatical case, gender, or number is used. 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. with eTranslation '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'). On the other part, some corrected segments were identical even though the source was different due to the use of articles in the English language (e.g. 'North Sea' and 'the North Sea' were both translated as 'Severno morje'). This occurred with both MT, but in some cases Google CT performed better, producing more exact matches. It was also better at capturing the same amount or length of answers as in the original. The answer of eTranslation for 'harvests of their Chinese tenants' was: 'čemer je dohodek od žetve kitajskih najemnikov', whereas Google CT captured only 'žetve njihovih kitajskih najemnikov'.

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?').

6.3 Qualitative analysis of predictions

By observing the individual cases of incorrect predictions, the main source of errors seems to stem from the grammatical and stylistic errors of the machine translation and occasionally its inability to convey the right meaning. The correct predictions are most likely 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.

In the examples provided, we can see that there are two types of errors that we looked at. The first is when there is a wrong answer, but a right prediction (in Table 8), and the second is the correct answer and

the wrong prediction (Table 9). Most of the time, the wrong answers and predictions occur with the eTranslation service, and improvement of Google CT and HT is visible from a few representative examples, but sometimes, when the questions are more complicated, even the Google CT and HT do not provide a prediction at all, while sometimes only HT provides the correct prediction.

Table 8: *Examples of correct predictions with wrong answers*

#	Dataset	Question	Answer	Prediction
1	ENG	How many of Warsaw's inhabitants spoke Polish in 1933?	833,500	833,500
	ET	Koliko prebivalcev Varšave je leta 1933 govorilo poljsko?	prebivalcev	833.500
	GCT	Koliko prebivalcev Varšave je leta 1933 govorilo poljsko?	833.500	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 je
	ET	Kdo je posnel „Walking in Fresno?"	Bob	Bob Gallion
	GCT	Kdo je posnel "Walking in Fresno"? Kdo je posnel »Walking in Fresno«?	Bob Gallion	Bob Gallion
	HT	Kdo je posnel "Walking in Fresno"? Kdo je posnel »Walking in Fresno«?	Bob Gallion	Bob Gallion

Note. ENG denotes the English dataset, ET one translated by the eTranslation service, GCT one translated by the Google Cloud translation service, and HT one translated by a human.

Table 9: *Examples of correct answers with wrong predictions. ENG denotes the English dataset, ET one translated by the eTranslation service, GCT one translated by the Google Cloud translation service, and HT one translated by a human*

#	Dataset	Question	Answer	Prediction
1	ENG	How many total acres is Woodward Park?	300 acres	300 acres
	ET	Koliko hektarjev je Woodward Park?	300 hektarjev	235 hektarjev
	GCT	Koliko skupno hektarjev obsega Woodward Park?	300 hektarjev	300
	HT	Koliko akrov skupaj obsega park Woodward?	300 akrov	300
2	ENG	How many miles, once completed, will the Lewis S. Eaton trail cover?	22 miles	22
	ET	Koliko kilometrov, ko bo konč ano, bo pokrivalo Lewis S. Eaton?	22 milj	(35 km
	GCT	Koliko milj bo pot Lewisa S. Eatona pokrivala, ko bo konč ana?	22 milj	22
	HT	Koliko milj bo, ko bo dokonč ana, dolga pot Lewisa S. Eatona?	22 milj	22

7 Conclusion

In this work, we presented a method for the machine translation of question answering datasets. To evaluate the method, we applied it to the SQuAD 2.0 dataset and used the results to train and test the following question answering models: Multilingual BERT, CroSloEngual, SloBERTa 2.0, XLM-RoBERTa, and RemBERT. In order to discern the impact of the quality of the translated data we performed the translation with two different translation services: eTranslation and state-of--the-art Google Cloud Translation. To evaluate the models using as close to real data as possible, we took a small subset of the original testing set and manually translated it to Slovene, which formed the basis for the performance comparisons.

The results show that using machine-translated data for evaluation led to notably worse results as compared to the human-translated data. 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 as available via the eTranslation 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 results show that the application of machine-translated data produced by our method leads to better results on smaller multilingual question answering models, as compared to fine-tuning them using the original, English, data. On the other hand, the results for larger models were mostly unaffected by the language of the dataset used to train them. The most notable benefit is the ability to fine-tune monolingual models, which would otherwise be unusable. Our experiments also show that this machine-translated data is not suitable for the purpose of testing the models. The impact of the quality of the translation service is minor and varies depending on the model.

The testing procedure could be improved by using a dataset that was already manually translated to Slovene, which would allow us to

benchmark our results against it as well. The experiment could also be expanded by including more datasets, such as Natural Questions (Kwiatkowski et al., 2019), and other models, such as Microsoft's mDeBERTaV3. Additionally, further effort could be dedicated to ascertaining the optimal parameters for fine-tuning the question answering models.

References

Arnejšek, M., & Unk, A. (2020). Multidimensional assessment of the eTranslation output for English–Slovene. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 383–392). Lisboa: European Association for Machine Translation. Retrieved from https://aclanthology.org/2020.eamt-1.41

Chung, H. W., Févry, T., Tsai, H., Johnson, M., & Ruder, S. (2021). Rethinking Embedding Coupling in Pre-trained Language Models. *International Conference on Learning Representations*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., . . ., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.747

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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* (Vol. 1, pp. 4171–4186). Minneapolis: Association for Computational Linguistics. doi: 10.18653/v1/N19-1423

European Commission. (2020). CEF Digital eTranslation. *CEF Digital eTranslation*.

Koehn, P., & Knowles, R. (2017). Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation* (pp. 28–39). Vancouver: Association for Computational Linguistics. doi: 10.18653/v1/W17-3204

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., . . ., & Petrov, S. (2019). Natural Questions: a Benchmark for Question

Answering Research. *Transactions of the Association of Computational Linguistics*.

Lample, G., & Conneau, A. (2019). Cross-lingual Language Model Pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *International Conference on Learning Representations.* Retrieved from https://openreview.net/forum?id=H1eA7AEtvS

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . ., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*.

Ljubešić, N., & Dobrovoljc, K. (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* (pp. 29–34). Florence: Association for Computational Linguistics. doi: 10.18653/v1/W19-3704

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., . . ., & Sagot, B. (2020). CamemBERT: a Tasty French Language Model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7203–7219). Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.645

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., . . ., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research, 21*, 1–67. Retrieved from http://jmlr.org/papers/v21/20-074.html

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. doi: 10.48550/ARXIV.1806.03822

Ulčar, M., & Robnik-Šikonja, M. (2020). Finest BERT and CroSloEngual BERT. *International Conference on Text, Speech, and Dialogue* (pp. 104–111).

Ulčar, M., & Robnik-Šikonja, M. (2021). SloBERTa: Slovene monolingual large pretrained masked language model. *In Proceedings of Data Mining and Data Warehousing, SiKDD*.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., . . ., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish. *arXiv preprint arXiv:1912.07076*.

Woods, W. A., & WOODS, W. A. (1977). Lunar rocks in natural English: Explorations in natural language question answering.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., . . ., & Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 483–498). Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems, 32*.

Zupanič, M., Zirkelbach, M., Šmajdek, U., & Jazbinšek, M. (2022). Preparing a corpus and a question answering system for Slovene. In D. Fišer & T. Erjavec (Eds.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (pp. 353–359). Ljubljana, Inštitut za novejšo zgodovino. Retrieved from https://nl.ijs.si/jtdh22/pdf/JTDH2022_Proceedings.pdf

Appendix A: Translation examples

Below are some examples of two machine translations and a human translation, where some specific differences, which occur more times, can be seen.

1. **Example of context segment (excerpt)**
    - 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.
    - eTranslation
        - 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.
    - Google Translate
        - Severni Kitajci so bili uvrščeni višje, južni Kitajci pa nižje, ker je južna Kitajska zdržala in se borila do zadnjega, preden je popustila.
    - Human 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.
2. **Examples of answerable question segment**
    - Original
    - Who was Al-Banna's assassination a retaliation for the prior assassination of?
    - What plants create most electric power?
    - eTranslation
    - Kdo je bil Al-Bannin umor maščevanja zaradi predhodnega umora?
    - Katere rastline ustvarjajo največ električne energije?
    - Google CT
    - Komu je bil atentat Al-Banne povračilo za prejšnji atentat?
    - Katere rastline proizvajajo največ električne energije?
    - Human translation
    - Al-Bannov umor je bil maščevanje za čigav predhodni umor?
    - Kateri obrati ustvarjajo največ električne energije?

3. **Example of impossible question segment**
    - Original
        - What Book of the Bible is knowledge of the law traced back to?
    - eTranslation
        - Do katere knjige Svetega pisma je znano pravo?
    - Google CT
        - Od katere svetopisemske knjige sega znanje o zakonu?
    - Human translation
        - V kateri svetopisemski knjigi že zasledimo poznavanje prava?

Appendix B: HTML Structure

```
<data class=0>
   <paragraph class=0>
      <context>The Normans (Norman: Nourmands; French: Normands; Latin:
      Normanni) were the people who in the 10th and 11th centuries gave
      their name to Normandy, a region in France. They were descended
      from Norse ("Norman" comes from "Norseman") raiders and pirates
      from Denmark, Iceland and Norway who, under their leader Rollo,
      agreed to swear fealty to King Charles III of West Francia. Through
      generations of assimilation and mixing with the native Frankish
      and Roman-Gaulish populations, their descendants would gradually
      merge with the Carolingian-based cultures of West Francia. The
      distinct cultural and ethnic identity of the Normans emerged ini-
      tially in the first half of the 10th century, and it continued to
      evolve over the succeeding centuries.</context>
   <qas class=0>
   <question>In what country is Normandy located?</question>
   <answer class=0>
   <text>France</text>
   </answer>
   </qas>
   <qas class=1>
   <question>When were the Normans in Normandy?</question>
   <answer class=0>
   <text>10th and 11th centuries</text>
      <in_context>The Normans (Norman: Nourmands; French: Normands; La-
      tin: Normanni) were the people who in the <b>10th and 11th cen-
      turies</b> gave their name to Normandy, a region in France. They
      were descended from Norse ("Norman" comes from "Norseman") raiders
      and pirates from Denmark, Iceland and Norway who, under their le-
      ader Rollo, agreed to swear fealty to King Charles III of West
      Francia. Through generations of assimilation and mixing with the
      native Frankish and Roman-Gaulish populations, their descendants
      would gradually merge with the Carolingian-based cultures of West
      Francia. The distinct cultural and ethnic identity of the Normans
      emerged initially in the first half of the 10th century, and it
      continued to evolve over the succeeding centuries.</in_context>
```

```
   </answer>
   <answer class=1>
   <text>in the 10th and 11th centuries</text>
      <in_context>The Normans (Norman: Nourmands; French: Normands; La-
      tin: Normanni) were the people who <b>in the 10th and 11th cen-
      turies</b> gave their name to Normandy, a region in France. They
      were descended from Norse ("Norman" comes from "Norseman") raiders
      and pirates from Denmark, Iceland and Norway who, under their le-
      ader Rollo, agreed to swear fealty to King Charles III of West
      Francia. Through generations of assimilation and mixing with the
      native Frankish and Roman-Gaulish populations, their descendants
      would gradually merge with the Carolingian-based cultures of West
      Francia. The distinct cultural and ethnic identity of the Normans
      emerged initially in the first half of the 10th century, and it
      continued to evolve over the succeeding centuries.</in_context>
   </answer>
   </qas>
   ...\\
   </paragraph>
   ...\\
</data>
```

Appendix C: Detailed results

Table 10: *Full comparison of the results of various models and their fine-tuning configurations on the human-translated subset of SQuAD 2.0 (N=285), and the subsets containing the same question from the original English dataset and the two machine-translated datasets*

Model Name	Fine-Tuning Dataset	Original [%]		eTranslation [%]		Google CT [%]		Human Transl. [%]	
		Exact	F1	Exact	F1	Exact	F1	Exact	F1
M-BERT	Original	76.1	78.4	42.8	48.2	53.0	55.9	55.1	58.2
	eTranslation	58.2	62.6	58.6	64.5	68.7	73.4	70.9	74.0
	Google CT	59.3	65.1	58.9	64.8	66.7	71.4	69.4	73.6
CroSloEngual BERT	Original	73.3	75.5	53.0	60.8	57.5	63.4	61.1	65.1
	eTranslation	59.6	63.1	51.6	58.8	58.2	64.5	60.0	63.6
	Google CT	55.8	58.8	60.7	66.5	61.8	66.6	70.2	73.6
SloBERTa 2.0	eTranslation	59.3	65.0	64.9	72.2	69.5	76.1	70.9	74.9
	Google CT	61.8	65.2	61.4	68.0	66.3	72.9	73.0	78.3
XLM-R Large	Original	83.5	85.5	61.4	69.1	70.9	75.8	**78.6**	**81.6**
	eTranslation	79.3	82.6	**66.0**	**73.1**	71.2	76.9	76.1	81.1
	Google CT	79.3	82.3	63.5	70.9	**71.2**	**77.4**	76.8	81.4
RemBERT	Original	**84.9**	**87.2**	64.2	71.4	69.1	74.3	74.0	79.5
	eTranslation	80.0	84.1	64.9	72.9	70.2	76.6	71.9	78.6
	Google CT	80.4	84.8	63.2	71.6	70.2	76.0	**77.9**	**81.5**

Note. Specific parameters used in fine-tuning are presented in Table 5.

Table 11: *Full comparison of the results of various models and their fine-tuning configurations on the English SQuAD 2.0 evaluation dataset and the two Slovene machine-translated SQuAD 2.0 evaluation datasets (N=11.680)*

Model Name	Fine-Tuning Dataset	Original [%]		eTranslation [%]		Google CT [%]	
		Exact	F1	Exact	F1	Exact	F1
M-BERT	Original	75.8	78.9	52.7	59.2	57.1	61.9
	eTranslation	63.7	68.2	61.7	68.3	65.7	70.7
	Google CT	64.4	68.9	61.4	67.9	66.9	71.3
CroSloEngual BERT	Original	72.9	76.3	56.2	63.5	61.5	66.8
	eTranslation	63.6	68.2	58.3	65.5	62.3	68.3
	Google CT	61.4	65.7	59.8	66.5	65.3	70.0
SloBERTa 2.0	eTranslation	60.6	64.7	66.6	73.7	71.4	76.8
	Google CT	63.9	66.9	65.6	72.8	72.3	77.0
XLM-R Large	Original	83.4	86.3	67.1	74.8	73.4	78.5
	eTranslation	79.0	83.0	**68.0**	**75.6**	72.3	78.3
	Google CT	80.9	84.4	**68.0**	**75.5**	**75.3**	**80.1**
RemBERT	Original	**84.5**	**87.5**	67.1	71.4	69.1	74.3
	eTranslation	79.1	83.9	66.8	72.9	70.2	76.6
	Google CT	80.1	84.5	67.0	71.6	70.2	76.0

Note. The English dataset only contains the questions pre-set in its Slovene counterpart. Specific parameters used in fine-tuning are presented in Table 5.

Prilagoditev angleškega korpusa in sistema za odgovarjanje na vprašanja za slovenski jezik

Pomanjkanje ustreznih podatkov za učenje je ena od ključnih težav pri razvoju slovenskih modelov za odgovarjanje na vprašanja (QA). Sodobna orodja za strojno prevajanje lahko to težavo rešijo, vendar pa se pri njihovi uporabi soočimo z novih izzivom: odgovori se morajo natančno ujemati z deli danega konteksta, kjer ta odgovor je, saj model odgovorov ne generira, temveč le išče. Kot rešitev predlagamo metodo, kjer odgovore prevajamo skupaj s kontekstom, kar poveča verjetnost, da bo odgovor preveden v enaki obliki. Učinkovitost te metode ocenjujemo na naboru podatkov SQuAD 2.0, prevedenem z uporabo storitev eTranslation in Google Cloud, kjer se z njeno uporabo delež neujemanj odgovora in konteksta zmanjša s 56 % na 7 %. Prevedene podatke nato ocenimo z uporabo različnih QA modelov, ki temeljijo na arhitekturi transformer, in preučimo razlike med podatkovnimi nizi in

konfiguracijami modelov. Da zagotovimo čim bolj realistične rezultate, modele testiramo na človeških prevodih majhnega deleža izvirne zbirke podatkov. Rezultati kažejo, da se glavne prednosti uporabe strojno prevedenih podatkov pokažejo pri natančnem prilagajanju (angl. fine-tuning) manjših večjezičnih modelov in enojezičnih modelov. Večjezični CroSloEngual BERT model je na primer dosegel 70,2 % točnih ujemanj pri testiranju na slovenskih podatkih v primerjavi s 73,3 % točnih ujemanj pri testiranju na angleških podatkih. Medtem ko so bili rezultati pri večjih modelih podobni, pri čemer je RemBERT dosegel 77,9 % točnih ujemanj na slovenskih podatkih v primerjavi z 81,1 % na angleških podatkih, so se ti obnesli podobno tudi pri natančnem prilagajanju na angleških podatkih, kar pomeni, da jih strojno prevedeni podatki niso bistveno izboljšali.

Ključne besede: sistemi za odgovarjanje na vprašanja, strojno prevajanje, večjezični modeli