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

Cross-Lingual Question Answering with Transformers

Matjaž Zupanič, Maj Zirkelbach, Uroš Šmajdek and Meta Jazbinšek

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 [1] dataset, which we fully translate using machine translation. We obtain a benchmarking dataset by manually post-editing the small subset of machine translations. We then compare these datasets by various transformer-based QA models, and observe the differences between the datasets and different model configurations. The results have shown that monolingual models perform best, even if the multilingual model was first fine-tuned on English language. 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.

Keywords

question answering, transformers, translation, multilingual models

Advisors: Slavko Žitnik, Špela Vintar

Introduction

A core goal in artificial intelligence is to build systems that can read 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) [2] and A Lite BERT (ALBERT) [3] 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,

In this work we present a method for a construct of a machine translated dataset from SQuAD 2.0 [1] and evaluate its quality using various modern QA models. Additionally, we benchmark its effectiveness by performing a manual post editing on a subset of the translated dataset and comparing the results.

The main contributions of our work are:

- a pipeline for reliable 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.

Related work

Early question answering systems, such as LUNAR [4], 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 perfor-

mance, the focus shifted from such rule-based system to using machine learning and statistical approaches, such as 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. [5] 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 [2], a transformer model introduced in 2019. Since then it has inspired many other transformed based models, for instance RoBERTa [6], ALBERT [3], and T5 [7] , xlm and XLNet [8].

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 [9], mT5 [10] and RemBERT [11]. 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 [12, 13]. As such Ulčar et al. [14] 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. [14].

In 2020 a Slovene monolingual RoBERTa-based model SloBERTa [15] was introduced. It was trained on 5 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.

Dataset

Stanford Question Answering Dataset (SQuAD 2.0) [1] is a reading comprehension dataset. It is based on a set of articles on Wikipedia covering different topics, where every question is a segment of text, or span, from the corresponding reading passage. This topics for example include historical, pharmaceutical, religious, as well as European Union texts. 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. It should be noted that the database SQuAD 2.0 is not entirely reliable. From the batch of 142 test question and answer groups, there were 14 occurrences where at least one of the given answers was not correct (i.e. Advanced Steam movement instead of pollution). Although it is a small percentage, we could not manually check the whole database.

Automatic Translation

To translate the dataset into Slovenian we used the eTranslation webservice [16]. 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 4MB 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. 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 library [17] was used. If a match was found we replaced the bad answer with the original text forming the lemma sequence in the context. Second 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.

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

Basic	Corrected	In-context	Corrected in-context
44%	66%	93%	94%

Post-Editing of Automatic Translation

Post-editing was done on random automatically translated excerpts. The provided excerpts included original paragraphs, questions and answers, as well as their automatic 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 which was generated from automatic 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 improvements, but on correcting the grammatical errors, wrong meanings and very unusual syntax. As mentioned above, the topics of original texts are diverse and very technical, covering domains such as religion, history, politics, mathematics and chemistry.

Out of 780 segments in the project, 30 were internal markers for context-questions-answers groups, which were not part of analysis. Each context had a varying number of questions and answers. The number of different segment types and of post-editing changes can be seen in Table 2.

Table 2. 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.

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 %

Post-Editing Analysis

Though the numbers seen in Table 2 are still not perfectly representative, since some corrections are more severe than others and somewhere there is a much greater number of them in one segment, we can see none of the paragraphs with context was without them, which is predictable, as the phrases are complicated and the texts include very specific terminology.

Here is one example of a sentence with more severe semantic errors:

1. 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. 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. 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, which is quite high because automatic translation provided incoherent results. Here the changes are also more notable, because they affect the overall understanding for potential users. Below are some examples of such questions:

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?

Automatic 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?

Human 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, even though the percentage of corrected ones is quite 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 and some more severe are semantic mistakes (i.e. plants translated as 'rastline', not 'naprave'), completely wrong answers (i.e. empty segment instead of 'Fermilab' or 'in' instead of '1.388'), some frequent ones were also the names of movements, books, projects or names (i.e. 'Bricks for Varšava' was left untranslated, 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 already. 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 (i.e. '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 English language (i.e. 'North Sea' and 'the North Sea' were both translated as 'Severno morje').

Models

In this section we give a brief overview of the five models we used in the evaluation. We also present the results of three massively multilingual models, xlm-R, M-BERT and RemBERT, on XTREME multilingual question answering benchmark in Table 3.

xlm-R

xlm-R (xlm-RoBERTa) [9] is a pre-trained cross-lingual language model based on xlm [18]. 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.

M-BERT

M-BERT (Multilingual Bert) [19] is a pre-trained cross-lingual language model as it's name suggest. It is based on BERT [2]. The pre-trained model is trained on 104 languages with large amount of data from Wikipedia, using a masked language modeling (MLM) objective. On Hugginface, there is only a base model with with 12 hidden transformer layers available, large model with 24 hidden transformer layers was not uploaded and we were not able to test it.

RemBERT

RemBERT [20] is a pre-trained model 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.

Table 3. Comparison of different multi-lingual question-answering models on XTREME benchmark [21].

Model	f1 metric
mBERT	53.8%
RemBERT	68.6%
XLM-R (large)	62.3%

SloBERTa

SloBERTa [15] 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 different tokenization model. Since the model requires a large dataset for training, it was trained on 5 combined datasets. It outperformed existing Slovene models.

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 [13].

Results

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.

All tests were performed on i5 10400f system with RTX 3070 GPU 8GB VRAM. For larger models we used RTX 3060 12GB. 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 evalaluation and were as follows:

- **Exact** - The fraction of predictions matched at least of one the correct answers exactly.
- **F1** - Which measures the average overlap between prediction and ground truth and is 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 on the non-translated SQuAD 2.0 and machine translated dataset can be seen in Table 5. The results on 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.

Table 4. Parameters used to fine-tune different configurations of the evaluated models. Language denotes which set was used for fine-tuning, specifically Slo represents the machine translated SQuAD 2.0 and Eng represents the original English SQuAD 2.0 dataset. The latter only contains the questions that are present in machine translated dataset. *B* denotes the number of batches used during fine-tuning, *MS* the maximum sequence length, *LR* the learning rate and *E* number of epochs.

Model Name	Language	B	MS	LR	E
xlmR-large	Eng	32	256	1e-5	3
xlmR-large	Slo	4	256	3e-5	3
xlmR-large	Eng & Slo	4	256	3e-5	3
M-BERT-base	Eng	8	320	3e-5	3
M-BERT-base	Slo	8	320	3e-5	3
M-BERT-base	Eng & Slo	8	320	3e-5	3
CroSloEngual BERT	Eng	4	256	1e-5	3
CroSloEngual BERT	Slo	4	256	1e-5	3
CroSloEngual BERT	Eng & Slo	4	256	1e-5	3
RemBERT	Eng	4	256	1e-5	3
SloBERTa 2.0	Slo	16	320	3e-5	3

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	
		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 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.

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%**

Discussion

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, which is also evidenced by its better performance on multilingual XTREME benchmark as shown in Table 3. While 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.

By comparing the machine and human translated results in Table 6, we can see that the human translation performs better across all models, suggesting that the automatic 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.

Qualitative Analysis

While there are many correct predictions of the answers in the automatically translated dataset, it is clear that a great number of predictions still does not answer the question correctly. This is because the automatic translation of the sentences

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	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 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.

#	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

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 from only a few representative examples in Table 7 and Table 8, improvement of some answers is already visible.

Conclusion

In this work we present a machine translated SQuAD 2.0 dataset and evaluate it on the following question answering (QA) models: xlmR-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 result 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 [16] 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. Memory and speed limits of the RTX 3070 GPU prevented us from fine-tuning the larger RemBERT dataset with comparable parameters to other models and in a decent time-frame. 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 [22], and different machine translation services.

References

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

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

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

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

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

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

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

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

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

[10] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

[11] Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821*, 2020.

[12] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*, 2019.

[13] Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*, 2019.

[14] Matej Ulčar and Marko Robnik-Šikonja. Finest bert and crosloengual bert. In *International Conference on Text, Speech, and Dialogue*, pages 104–111. Springer, 2020.

[15] Matej Ulčar and Marko Robnik-Šikonja. Sloberta: Slovene monolingual large pretrained masked language model. 2021.

[16] CEF Digital etranslation. https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation. Accessed: 2022-05-23.

[17] Nikola Ljubešić and Kaja Dobrovoljc. 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, August 2019. Association for Computational Linguistics.

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

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

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

[21] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080, 2020.

[22] 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. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.