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

Cross-lingual question answering with computer

Matjaž Zupanič, Maj Zirkelbach, Uroš Šmajdek and Meta Jazbinšek

Abstract

In this paper we discuss automatic cross-lingual question answering based on machine learning. We are only focused on English and Slovene language.

Keywords

cross-lingual, question answering, machine learning, automated ...

Advisors: Slavko Žitnik

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) [1] and A Lite BERT (ALBERT) [2] 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, especially when translating between different language groups.

In our project we wish to evaluate various QA models on different types of corpora; original English variants, those machine translated into Slovene and those that were manually checked by a human after machine translation.

Related work

Mention:

- BERT
- RoBERTa
- SloBERTa
- RemBERT
- XLnet
- ERNIE-M?
- XLM
- XLM-R
- T5
- mT5
- Slovene SuperGLUE

(Other) related datasets:

- HotpotQA
- bAbI
- TriviaQA
- WikiQA
- XTREME (benchmark, 3 QA subsets)

Methods

Dataset(s)

Stanford Question Answering Dataset (SQuAD 2.0) [3] is a reading comprehension dataset. It is based on a set of articles on Wikipedia, where every question is a segment of text, or span, from the corresponding reading passage. Otherwise question might be unanswerable. It consists over 100.000 questions-answers extracted from over 500 articles. Reason to use Squad 2.0 over 1.0 is that it consists twice as much data and contains unanswerable questions.

Natural Questions (Alternate) is Google's dataset for question answering [4]. It contains questions from real users

and it requires QA systems to read and comprehend an entire Wikipedia article which may or may not contain the answer to confuse system. So it is one of the most realistic and challenging sets for question answering. It consists over 307.000 training examples, over 7.830 development examples, and over 7.842 test examples.

Translation

To translate the dataset into Slovenian we use the eTranslation webservice [5]. Due to the webservice 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 curpus 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 [6] 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.

Models

XLM-R

XLM-R (XLM-RoBERTa) [7] is a pretrained cross-lingual language model based on XLM [8]. 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.

mBERT

MBERT (Multilingual Bert) [9] is a pretrained cross-lingual language model as it's name suggest. It is based on BERT [1]. Pretrained model is trained on 104 languages with large amount of data from Wikipedia, using a masked language modeling (MLM) objective. There is only base version available.

mT5

Multilingual T5 (mT5) [10] is a massively multilingual pretrained text-to-text transformer model, trained following a similar recipe as T5 [11]. It is pretrained on the mC4 [12] corpus, covering 101 languages.

RemBERT

RemBERT [13] is a pretrained 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 1. Comparison of different multi-lingual question-answering models on XTREME benchmark [14].

Model	f1 metric
mBERT	53.8%
RemBERT	68.6%
XLM-R (large)	62.3%
mT5	73.6%

SloBERTa

SloBERTa [15] is Slovene monolingual large pretrained masked language model. It is based on RoBERTa model. Since model requires large dataset for training, it was trained on 5 combined datasets. It outperformed existing Slovene models.

Results

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 2.

Table 2. 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%

Question Answering

To evaluate the performance of mBERT and XLN-R models we first evaluated their performance on the non-translated SQuAD 2.0 dataset, in order to obtain the reference data. This results were then compared to the evaluations done on the translated SQuAD 2.0 dataset, with different degrees of fine-tuning. All tests were performed on i5 10400f system with RTX 3070 GPU 8GB VRAM. The results on non-translated SQuAD 2.0 dataset are visible on table 3 and the results on translated SQuAD 2.0 on table 4.

TODO

- Compare manually translated subset
- Evaluate mT5, RemBERT and SloBERTa??
- qualitative evaluation

Table 3. Performance metrics of non-fine tuned and fine-tuned xlm-roberta-base model on Squad2.0 dataset. The parameters used during fine-tuning and evaluation were: Batch size 8, max. sequence length 320, learning rate 3e-5, doc stride 128.

Epochs	exact metric	f1 metric
-	40.0%	40.1%
3	76.7%	79.9%
4	76.8%	80.1%

Table 4. Performance metrics of different models, evaluated on the translated Squad2.0 dataset. FT denotes that the model was fine-tuned. All fine-tuning was performed with 3 epochs.

Model name	exact metric	f1 metric
bert-base-multilingual-cased	5.4%	0.0%
bert-base-multilingual-cased FT	60.4%	67.0%
xlm-roberta-base, no tuning	39.2%	39.7%
xlm-roberta-base FT	61.6%	68.6%
xlm-roberta-large	64.3%	72.3%

References

[1] 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.

[2] 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.

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

[4] 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.

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

[6] 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.

[7] 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.

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

[9] 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.

[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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics.

[11] 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.

[12] 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. *arXiv e-prints*, 2019.

[13] 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.

[14] 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.

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