University of Ljubljana
Faculty of Computer and Information Science

FRI Natural language processing course 2022
Project report

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

We found several large datasets, many of which are generally recognised as benchmarks for Question Answering tasks. First we give a quick overview with links to dataset webpages, followed by a more in-depth description of each dataset:

- SQuAD 1.0 [3]
- SQuAD 2.0 [1] [4]
- Natural Questions by Google AI [2] [5]
- SuperGLUE [3] [6]
- Slovene SuperGLUE Benchmark [4] [7]

Stanford Question Answering Dataset (SQuAD 2.0) [4] 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 is Google's dataset for question answering [5]. 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.

SuperGlue [6] is a refined version of Glue benchmark tool consisting different benchmarks. It is comprised of 8 corpora (BoolQ, CB, COPA, MultiRC, ReCoRD, RTE, WiC,

[1] https://rajpurkar.github.io/SQuAD-explorer/
[2] https://ai.google.com/research/NaturalQuestions
[3] https://super.gluebenchmark.com/tasks
[4] https://www.clarin.si/repository/xmlui/handle/11356/1380

WSC), with 4 different types of tasks. One of those is a part for measuring question answering performance. Compared to other described datasets, it is much smaller , containing 9427 labeled training examples, 3270 labeled development examples, 3245 unlabeled test examples. Totaling 15942 examples.

Slovene SuperGlue [7] is Slovenian translation of SuperGlue. Some of the translations are translated by Google Machine Translation service (BoolQ, CB, COPA, MultiRC, and RTE), while others are translated by human (COPA and WSC completely, and BoolQ, CB, MultiRC, ReCoRD, and RTE in different ratios).

Additionally we give a brief overview of several widely recognised language models which are all able to perform question answering tasks. First we give a quick overview with links to the model repositories, followed by a more in-depth description of each dataset:

- BERT [5] [1]
- RoBERTa [6] [8]
- SloBERTa [7] [9]
- ALBERT [2]
- CORA [8] [10]
- Xlnet [9] [11]

BERT [1] is a Transformer based machine learning technique. It is an attention mechanism that learns contextual relations between words in a text. BERT has become a baseline for natural language processing (NLP), toping over 150 research publications, that modify or extend algorithm to perform better.Even Google uses BERT for it's search engines. In the following toturial authors used Bert Cross lingual question answering with DeepPavlov [12, 13], that outperformed human performance on SQuaD 2.0. DeepPavlov is a conversational artificial intelligence framework that contains all the components required for building chatbots, developed for TensorFlow and Keras.

M-BERT is a multilingual version of BERT that support 104 languages. Model allows to perform zero-shot transfer from source language to target language. For source language we usually use English, since it has largest datasets.

RoBERTa [8] is better, optimized method based on BERT. It modifies key hyperparameters and it is trained on larger dataset. It is also a part of Facebook's ongoing commitment to develop the state-of-the-art algorithm in self-supervised systems that can be developed with less reliance on time and resource-intensive data labeling.

SloBERTa [9] 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.

CORA [10] Cross-lingual Open-Retrieval Answer Generation can answer questions in many languages, even for ones without language-specific annotated data or knowledge sources. CORA answers directly in given language without any translation needed.

XLnet [11] is Generalized Autoregressive Pretraining for Language Understanding. Unlike BERT, which is based on bidirectional context and denoising autoencoding, the XLnet is based on autoregressive language modeling. XLNet authors claim that it outperforms BERT on 20 tasks, often by a large margin, including question answering.

[5] https://github.com/google-research/bert
[6] https://github.com/pytorch/fairseq/tree/main/examples/roberta
[7] https://www.clarin.si/repository/xmlui/handle/11356/1397
[8] https://github.com/AkariAsai/CORA
[9] https://github.com/zihangdai/xlnet

Results

Discussion

Acknowledgments

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, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. 2016.

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

[5] 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.

[6] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

[7] Aleš Žagar and Marko Robnik-Šikonja. Slovene superglue benchmark: Translation and evaluation. *arXiv preprint arXiv:2202.04994*, 2022.

[8] 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.

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

[10] Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. One question answering model for many languages with cross-lingual dense passage retrieval. *Advances in Neural Information Processing Systems*, 34, 2021.

[11] 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.

[12] Vasily Konovalov. Bert-based cross-lingual question answering with deeppavlov. https://towardsdatascience.com/bert-based-cross-lingual-question-answering-with-deeppavlov-704242c2ac6f, 2019. Accessed: 22.03.2022.

[13] Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. DeepPavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, Melbourne, Australia, July 2018. Association for Computational Linguistics.