Context-Based Question-Answering System for the Ukrainian Language

Serhii Tiutiunnyk and Vsevolod Dyomkin

Ukrainian Catholic University
Faculty of Applied Sciences, Lviv, Ukraine
tiutiunnyk@ucu.edu.ua, vseloved@gmail.com

Abstract. We introduce a context-based question answering model for the Ukrainian language based on Wikipedia articles using Bidirectional Encoder Representations from Transformers (BERT) [1] model which takes a context (Wikipedia article) and a question to the context. The result of the model is an answer to the question. The model consists of two parts. The first one is a pre-trained multilingual BERT model which are trained on the top-100 the most popular languages on Wikipedia articles. The second part is the fine-tuned model, which is trained on the data set of questions and answers to the Wikipedia articles. The training and validation data is Stanford Question Answering Dataset (SQuAD) [2].There is no any question answering datasets for the Ukrainian language. The plan is to build an appropriate dataset with machine translate and use it for the fine-tuning training stage and compare the result with models which were fine-tuned on the other languages. The next experiment is to train a model on the Slavic languages dataset before fine-tuning on the Ukrainian language and compare the results.

Keywords: Context-based Question Answering · Bidirectional Encoder · Representations from Transformers · multilingual BERT · fine-tuning · Generative Pre-trained Transformer · Stanford Question Answering Dataset

1 Introduction and Motivation

1.1 Problem Importance

Nowadays, it becomes more challenging to stay in the context of an expert area without handling huge volumes of data. Textual information grows exponentially together with video, audio, photo, and other types of data. Therefore, a model, which answers a question, is significant. It can be used for building chat-bots, automatic quiz generation. Finally, it helps to handle text documents and retrieve the necessary information much faster. For example, it is useful for companies with a massive base of inner instructions. Employees can retrieve the required data based on the scope of the documents.

In: Proceedings of the 1st Masters Symposium on Advances in Data Mining, Machine Learning, and Computer Vision (MS-AMLV 2019), Lviv, Ukraine, November 15-16, 2019, pp. 81–88

Along with it, the question answering system might be beneficial for layers, medical workers, and other specific professions.

1.2 General Formulation of the Problem

Question answering task is one of the classical problems in natural language processing (NLP). At the input for content-based question-answering model has a context and a question. As a context, we can take an article, a document, an essay, a paper, or any other piece of textual information. In this project, we will use articles from Wikipedia. A question is a natural human language question. Articles and questions are in the Ukrainian language. The result of the model is a phrase from the context, which contains the answer to the question.

2 Review of Related Work

Despite the importance of the problem, it is not appropriately solved for the Ukrainian language yet. There was no public result for the Ukrainian language found except some multilingual models like BERT [1].

2.1 Classical Methods

Let us start a review of existed methods from the classical approaches. Under the term classical, we mean methods, which use well-known strategies without artificial neural network models. There are unsupervised and supervised methods. Unsupervised approaches are based on word embedding [3] distances and word frequencies. Supervised methods use labeled dataset for training (logistic regression, support vector machine, etc.). Also, we can attribute logic-based methods (for example, Machine Comprehension Using Commonsense Knowledge [16]) to the set of classical methods. Such methods are used to solve question-answering task because logical representations yield more abstract concepts, such as temporal or logical relations. This is very useful for learning a type of commonsense knowledge.

Unsupervised Methods. Two different approaches are distinguished within this category of methods – based on measuring Euclidean distance between sentences and counting word and phrase frequencies.

Euclidean distance between sentences. The first traditional method we came across during reviewing of related works is finding the minimal Euclidean distance between question and sentences from the context [4]. The idea of this approach is to find an average vector of words for each sentence. The answer to the question is the closest sentence from the context to the question according to Euclidean distance. It is possible to specify the answer by splitting the sentence into phrases, but it is an additional task, which will decrease the accuracy of the method. One more drawback of the described method is relying on the quality of word embeddings. Also, this method does not take into account a dependency between the words in the sentence.

Word and phrase frequency. It is possible to use n-gram approach [5] for generating an answer. The question is parsed into the dependency tree and rebuilt into a narrative sentence with missing the target word or phrase. The missed phrase is filling by n-gram model. An artificial neural network model can replace the n-gram model. It will be discussed below. The drawback of this approach is a low accuracy of dependency parser models and relying on the phrase frequency in a relatively small volume of text.

Supervised Methods. This category of methods often use logistic regression and support vector machine approaches. Supervised traditional methods are described in [4]. The author uses the SQuAD dataset mentioned above for learning. Sentences from the context are split into the sentences and added to a binary vector. The target sentence is marked as 1 and all other items are 0. After that, multinomial logistic regression [6] is being trained by the labeled data or support vector machine [7]. One of the advantages of this approach is the ability to add some features to the model (dependency between the words, term frequency (TF), inverse document frequency (IDF) [8], etc.). A term frequency is a feature, which increases the weight of frequent words and inverse document frequency wise verse decreases the weight of widespread words.

Pros and Cons. The advantages of the classical approaches are simplicity and high transparency of the models. Along with it, the model performance on artificial samples is not good enough (near 70% accuracy on the SQuAD validation set). The result will be worse with increasing size of the context or setting a goal to retrieve a more specific answer (a phrase instead of a sentence). Moreover, the results for the Ukrainian language are even worse than the English language. It happens due to higher grammar complexity of the Ukrainian language, fewer text corpora, the presence of word cases and other language specifics.

2.2 Artificial Neural Network Models

In this part, we will review supervised and unsupervised cases for each main model.

Long Short-Term Memory Model. Long short-term memory model (LSTM) [9] is a recurrent neural network architecture, which allows building sequence-to-sequence models. Also, the input and output vector sizes are not fixed. As an input, LSTM model takes a context and a question and returns a word scores from the context. To connect a vector for context and a vector for a question, we add an attention layer. It is a crucial part of the question answering system based on LSTM model. Attention layer is a dot product of context and question output vectors. After that, the result of the dot product converts into the probability of being an answer to the question. The approach mentioned above is described in the paper dedicated to Bidirectional Attention Flow (BAF) [10].

Generative pre-trained transformer (GPT). There is a second version of this model called GPT-2 [11]. GPT-2 is one of the State-of-the-Art models in language modeling tasks. This model was trained on the Wikipedia articles and internet pages to make the style of generated text more various. This model can only generate the next word based on the previous text. Hence, to make it answer the question, we have to rephrase questions sentence into a narrative sentence with a skipped phrase for the answer. GPT-2 will generate the answer. The peculiarity of this model is the absence of the context. On the one hand, it can be an advantage if there is no specific data to retrieve the answer. On the other hand, the accuracy will be low for the tasks from special areas (law, medicine, etc.), as the model was not trained on data from the corresponded areas. Anyway, GPT-2 cannot be applied to the Ukrainian language, as it is trained only on English texts. Along with it, training the model from scratch or even pre-training on Ukrainian corpora requires a lot of resources and time.

2.3 Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) [1] is a transformer-based neural which shows state-of-the-art results in a wide variety of NLP tasks provided by Google researchers. Multilingual BERT model was built for top-100 of the most popular languages used in Wikipedia and it can be used for a hundred languages out-of-the-box. BERT model training process consists of two stages. The first stage is pre-training on the text corpora for language modeling task. The second stage is fine-tuning on the question-answer datasets. The first stage requires substantial computational resource. Fine-tuning, however, can be performed even on a single graphics processing unit (GPU).

Multilingual BERT results. There are several modifications of BERT multilingual models, which differ by the fine-tuning process. There are BERT models fine-tuned on a translated dataset, original dataset (English), cased (use original word case) and uncased (all words are lowercased). Table 1 shows the result of BERT modifications on Cross-lingual Natural Language Inference (XNLI) [13] dataset (translated datasets).

Table 1. BERT multilingual model performance

Model	English	Chinese	Spanish	German	Arabic	Urdu
BERT - Translate Train Cased	**81.9**	**76.6**	**77.8**	**75.9**	**70.7**	61.6
BERT - Translate Train Uncased	81.4	74.2	77.3	75.2	70.5	61.7
BERT - Translate Test Uncased	81.4	70.1	74.9	74.4	70.4	**62.1**
BERT - Zero Shot Uncased	81.4	63.8	74.3	70.5	62.1	58.3

As we can see, the best result for the vast majority of languages is provided by model pre-trained on the translated cased dataset. One more important thing is that the translated cased BERT model performs better for non-Latin alphabets languages [12].

3 Research Hypotheses and Problem

3.1 Hypotheses

Accuracy Hypothesis. The main objective of this project is to build a question-answer model for the Ukrainian language, which shows accuracy near results shown in Table 1 on the well-known benchmarks. The first hypothesis says it is possible to achieve an efficiency near 70-80%, which is close to results for the other languages provided by Google researchers.

Model Comparison Hypothesis. One more goal is to compare different approaches for pre-training. Some datasets have human translated data into the Russian and other Slavic languages. It seems that fine-tuning model on Slavic languages datasets and then fine-tuning on the turned into Ukrainian language dataset might improve performance for the Ukrainian language comparing with direct fine-tuning on the Ukrainian language dataset. So, the next task of this project is to confirm or deny this hypothesis.

3.2 Problems

Translation Problems. To achieve the project goals mentioned above, we need to find an appropriate machine translator to create the dataset in the Ukrainian language, build different model pipelines, and compare results. Furthermore, it might require a human translated small dataset in the Ukrainian language to verify the models.

Articles Retrieval Problems. Besides, the project needs to retrieve Wikipedia articles in the Ukrainian language. There are articles in the datasets which exist in the English Wikipedia and are absent in the Ukrainian part. Hence, we have to detect such items and exclude them from the datasets. Moreover, Wikipedia provides articles in the Extensible Markup Language (XML) format, which must be converted into the human-readable text.

4 Envisioned Approach

4.1 Dataset Generation

The very first task is to generate Ukrainian language dataset from the existing datasets (SQuAD [2]) by machine translator. There is a subtask related to the machine translation. It is comparison and checking the quality of the translation. The quality of the translated dataset directly affects the quality of the model. Translation quality can be checked by reverse translation. If the difference between the original text and the text after the forward and the backward translation is small enough, it indicates high quality of the translator.

4.2 Data Storing

Generated datasets and retrieved articles from Ukrainian Wikipedia are stored in the database to make access to the data more convenient. As the data size is bigger than read-only memory capacity, we will need to split and read data partially.

4.3 Models Pipeline

The base pre-trained model is multilingual BERT model. Then it is fine-tuned on the different datasets and variations. The first model will be fine-tuned on the translated training datasets (SQuAD). The accuracy of the model is calculated on the test sets of the corresponded datasets. The next model is fine-tuned on the human-translated datasets for Slavic languages. After that, the model will be fine-tuned on the machine-translated dataset for the Ukrainian language. Combinations on the fine-tuning stage produce different models, which are being compared on the test sets and the human-translated Ukrainian language dataset created manually.

5 Research Methodology and Plan

5.1 Methodological Approach

One can distinguish three methodological approaches [14]:

- Quantitative methods are appropriate for measuring, ranking, comparing, etc.
- Qualitative methods are best to measure describing, interpreting, contextualizing. Very often, it is related to the textual results.
- Mixed methods, which combine a numerical measurement and exploration.

On the one hand, quantitative methods are the best for comparison fine-tuned models between each other and with state-of-the-art models for the English language.

There will be applied the F1 [15] score and precision to get a quantitative measure and two types of matching. The first one is exact matching answers, and the second one will check if the original answer is present in the model answer.

On the other hand, sometimes answer to the question is not precisely equal to the expected value, but the meaning is correct. That is why mixed methods are the most appropriate for question-answering model evaluation. On this stage, the question-answer system may require a subsystem (additional artificial neural network or a simple set of rules) which decides if the answer is correct even if the model response is not equal to the labeled value.

Along with it, as it was mentioned above, the last stage of model evaluation is a human-translated test set in the Ukrainian language, which allows providing a qualitative measurement.

5.2 Plan for the Research

Table 2 shows a plan for the research.

Table 2. Timeline for the research

Milestone	**Start Date**	**End Date**
Coordination of the direction of the thesis	Aug 2019	Aug 2019
Review of past and current related work	Aug 2019	Oct 2019
Thesis proposal	Aug 2019	Sep 2019
Comparison machine translators	Sep 2019	Oct 2019

Milestone	Start Date	End Date
Translation datasets and building database of articles	Sep 2019	Oct 2019
Building baseline model	Oct 2019	Oct 2019
Building advanced fine-tuned models	Oct 2019	Dec 2019
Model evaluation	Nov 2019	Dec 2019
Writing master thesis	Oct 2019	Jan 2020
Submission of thesis for final review	-	8 Jan 2020
Master Thesis Defense	-	End of Jan 2020

6 Conclusive Remarks and Outlook

The most valuable thing from the potential results of the project is a high performance context-based question-answering model for the Ukrainian language. After the completion of this work, we will know how to build question-answering systems for the Ukrainian language. Further, these methods can be applied to the other Slavic languages or languages with very complicated grammar, peculiar properties, or non-Latin characters.

Translated datasets will be reusable for the other researches and projects and can be taken as a start point for the human translation process.

There is a point in the research plan where hypothesis might fail, and research must start from scratch. It is a hypothesis about building a high-performed question-answering model based on fine-tuning on the machine-translated datasets. This approach showed good results for English, Spanish, German, and Arabic languages. Along with it, the efficiency for the Urdu language is significantly worse than for the languages mentioned earlier (see Table 1).

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
2. Stanford Question Answering Dataset. https://rajpurkar.github.io/SQuADexplorer/
3. Al-Rfou, R., Perozzi, B., Skiena, S.: Polyglot: Distributed word representations for multilingual NLP. arXiv preprint arXiv:1307.1662 (2013)
4. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)
5. Sperandei, S., Understanding logistic regression analysis. Biochemia Medica **24**(1), 12–18 (2014). doi: 10.11613/BM.2014.003
6. Kwok, J.T.Y.: Automated text categorization using support vector machine. In: 5th International Conference on Neural Information Processing, pp. 347–351 (1998)
7. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York (1983)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation, **9**(8), 1735–1780 (1997)

9. Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603 (2016)
10. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8) (2019)
11. Google AI Research. https://github.com/google-research/bert
12. Cross-lingual Natural Language Inference dataset. https://github.com/facebookresearch/XNLI
13. Newman, I., Benz, C.R., Ridenour, C.S.: Qualitative-Quantitative Research Methodology: Exploring the Interactive Continuum. SIU Press, Carbondale and Edwardsville (1998)
14. Goutte, C., Gaussier, E.: A probabilistic interpretation of precision, recall, and F-score, with implication for evaluation. In: Losada D.E., Fernández-Luna J.M. (eds.) ECIR 2005. LNCS, vol. 3408, pp. 345–359. Springer, Berlin, Heidelberg (2005)
15. Apidianaki, M., Mohammad, S., May, J., Shutova, E., Bethard, S., Carpuat,M.: Proceedings of the 12th International Workshop on Semantic Evaluation. Association for Computational Linguistics (2018)
16. Ostermann, S., Roth, M., Modi, A., Thater, S., Pinkal, M.: SemEval 2018 Task 11: Machine comprehension using commonsense knowledge. In: 12th International Workshop on Semantic Evaluation, pp. 747–757. Association for Computational Linguistics (2018)