IRSTI 83.03.03

N. Ismagulov [1]

[1]Suleyman Demirel University, Kaskelen, Kazakhstan

QUESTION ANSWERING SYSTEM UPON UNIFIED LANGUAGE MODEL AND EVALUATING PERFORMANCE OF DATASETS

Abstract. Present days require automation and optimization in simple but urgent tasks. It is granted to use opportunities of technologies and science in order to work efficiently and to stay productive. In this paper, I seek to understand opportunities and drawbacks of the publicly available datasets, such as SQuAD, TriviaQA, Natural Questions (NQ), QuAC, NewsQA. It is vital to choose a suitable dataset in order to create a system with better performance. Specifically, the paper proposes an automatic question creating system that uses state-of-the-art Natural Language Processing (NLP) - Unified Language Model (UniLM). The question generating algorithm was verified using best datasets, and it has shown noteworthy results - questions generated were logical and correct. This study is important for teachers, teacher assistants, to save time writing test questions and spend it for more important duties.

Keywords: NLP, SQuAD, dataset, UniLM, education, wikipedia.

***

Аңдатпа. Қазіргі уақытта қарапайым, бірақ шұғыл тапсырмаларда автоматтандыру мен оңтайландыру қажет. Тиімді жұмыс істеу және өнімді болу үшін технологиялар мен ғылымның мүмкіндіктерін пайдалануға беріледі. Бұл мақалада мен SQuAD, TriviaQA, Natural Questions (NQ), QuAC, NewsQA сияқты жалпыға қолжетімді деректер жиынының мүмкіндіктері мен кемшіліктерін түсінуге тырысамын. Жақсырақ жұмыс істейтін жүйені құру үшін сәйкес деректер жинағын таңдау өте маңызды. Атап айтқанда, мақалада ең заманауи табиғи тілді өңдеу (NLP) - Бірыңғай тіл үлгісі (UniLM) қолданатын автоматты сұрақ жасау жүйесі ұсынылады. Сұрақтарды құру алгоритмі ең жақсы деректер жинақтары арқылы тексерілді және ол назар аударарлық нәтижелер көрсетті - құрастырылған сұрақтар логикалық және дұрыс болды. Бұл зерттеу мұғалімдерге, мұғалімдердің көмекшілеріне тест сұрақтарын жазу уақытын үнемдеу және оны маңыздырақ жұмыстарға жұмсау үшін маңызды.

Түйін сөздер: NLP, SQuAD, деректер жинағы, UniLM, википедия, білім беру.

***

Аннотация. Нынешние дни требуют автоматизации и оптимизации в простых, но срочных задачах. Разрешено использовать возможности технологий и науки, чтобы работать эффективно и оставаться продуктивным. В этой статье я стремлюсь понять возможности и недостатки общедоступных наборов данных, таких как SQuAD, TriviaQA, Natural Questions (NQ), QuAC, NewsQA. Крайне важно выбрать подходящий набор данных, чтобы создать систему с более высокой производительностью. В частности, в документе предлагается система автоматического создания вопросов, использующая современную обработку естественного языка (NLP) — Unified Language Model (UniLM). Алгоритм генерации вопросов был проверен с использованием лучших наборов данных, и он показал заслуживающие внимания результаты - сгенерированные вопросы были логичными и правильными. Это исследование важно для учителей, помощников учителей, чтобы сэкономить время на написании контрольных вопросов и потратить ее на более важные обязанности.

Ключевые слова: НЛП, SQuAD, набор данных, UniLM, википедия, образование.

1. Introduction

One of the improving fields of computer science is Artificial intelligence (AI). This is the process of machines simulating human intellect. There are popular applications such as Natural Language Processing (NLP), speech recognition and others. [1] Cognitive skills in which AI focuses on are learning, reasoning and self-correction. Learning aspect is responsible for obtaining data and creating algorithms. Reasoning aspect of AI programming iterates over algorithms to reach desired output. Self-correction is a self-explanatory aspect, which is about tuning over the algorithms for the accurate results. One of the popular examples of application of AI is GPT-3, which is a time series model that uses deep learning to create a human-like text. It allows people to communicate with machines in English. That means by simply describing what one wants to do one can get: code (websites, machine learning models, design), completed sentences, layout, simple reasoning.

There is also an increasing interest in investigating what models learn from datasets. Previous research in question answering has discovered that state-of-the-art models can perform well with partial input, rely too little on essential terms, and rely too much on basic heuristics. Experiments on SQuAD models in particular have revealed that they are subject to adversarial assaults and are not resistant to paraphrases.

During the course of the research I will be analyzing multiple datasets available publicly and state-of-the-art papers on both Natural Language Understanding and Natural Language Generation. Generally speaking, Natural Language Generation and Natural Language Understanding are subsections of Natural

Language Processing [2]: NLU reads and analyzes the text for grammar, style, and word-in-context meaning for intent and entities, whereas NLG creates a text from structured data.[3]

I propose automatic test generation based on reading material, an application that uses state-of-the-art Natural Language Processing (NLP) technologies to generate the questions based on the provided best performing, suitable for UniLM dataset material. The proposed work solves this problem by automating this process and converting key-term assignment into the test assignment that serves the same goal: Make students go through material and check how well they did it. Proposed solution will work on generating open questions. The concept of algorithm is to create questions with answers from input text material.

2. Literature review

2.1 "Unified Language Model Pre-training for Natural Language Understanding and Generation": Li Dong, Wenhui Wang, Nan Yang, Furu Wei, Jianfeng Gao, Ming Zhou, Xiaodong Liu Yu Wang, Hsiao-Wuen Hon.

This study proposes another UNIfied pre-prepared Language Model (UNILM) that may be modified for ordinary language comprehension as well as age-related tasks. In recent years, a new approach known as transformers has evolved in NLP and has supplanted all prior deep learning models, including RNN. Even though there are many NLG models like T5, BART, GPT developed by IT giants like Google, Facebook, OpenAI respectively, I chose the UniLM model developed by Microsoft. UniLM is pre-trained using unidirectional, bidirectional, and sequence-to-sequence prediction modeling tasks.The Unified Modeling is accomplished by utilizing a common Transformer organization and using explicit self-consideration covers to control what setting the forecast conditions on. UniLM contrasts well and BERT on the GLUE benchmark, and the SQuAD 2.0 and CoQA question noting assignments. Additionally, UniLM accomplishes new cutting edge results on five common language age datasets. After thorough analysis of the alternatives, UniLM fits the best for the project in terms of size, performance and speed, which will be demonstrated in research.

2.2 *"Language Models are Unsupervised Multitask Learners" : Alec Radford, David Luan, Jeffrey Wu, Rewon Child, Dario Amodei, Ilya Sutskever.*

Natural Language Processing tasks are normally drawn closer with supervised learning on task specific datasets. In this paper, authors exhibit that language models start to become familiar with these undertakings with no unequivocal oversight when prepared on another dataset of millions of web pages called WebText. The limit of the language model is basic to the accomplishment of zero-shot errand move and expanding it improves execution in a log-straight design across errands. Our biggest model, GPT-2, is a 1.5B boundary Transformer that accomplishes best in class results on 7 out of 8 tried languages displaying datasets in a zero-shot setting yet at the same time underfits WebText. Tests from the model mirror these upgrades and contain lucid passages of text. These discoveries recommend a promising way towards building language preparing frameworks which figure out how to perform errands from their normally happening exhibitions.

2.3 *"Attention Is All You Need": Ashish Vaswani, Niki Parmar, Noam Shazeer, Jakob Uszkoreit, Aidan N. Gomez, Llion Jones, Lukasz Kaiser, Illia Polosukhin.*

This paper proposes the Transformer, a model architecture that avoids recurrence and rather completely depends on a system of attention to draw global dependencies between input and output.

The Transformer allows significantly more parallelization and a new state of the art can be achieved in level of translation after being trained on eight P100 GPUs for as little as twelve hours.

3. Datasets

Table 1: Training and testing set sizes analytics of datasets

Dataset	*Training data amount*	*Testing data amount*
SQuAD	*130 319*	*11 873*
TriviaQA	*110 647*	*14 229*
NQ	*110 857*	*3 368*
QuAC	*83 568*	*7 354*
NewsQA	*101 707*	*5 666*

SQuAD, TriviaQA, Natural Questions, QuAC, and NewsQA are the datasets to compare. Each dataset is described here.

SQuAD 2.0 is made up of 150K question-and-answer pairs from Wikipedia articles. Crowd workers wrote questions regarding a Wikipedia passage and highlighted the answers to generate SQuAD 1.1. SQuAD 2.0 now adds an extra 50K plausible but unanswered queries posed by crowd workers. [4]

TriviaQA has 95K question-answer pairs gathered from trivia websites. The questions were developed by trivia lovers, and the evidentiary papers were obtained retroactively by the writers. I use the TriviaQA variation in which the documents are Wikipedia articles. [5]

Natural Questions (NQ) is a database of 300,000 questions culled from Google search records. A crowd worker found a lengthy and short response on a Wikipedia page for each topic. [6]

QuAC has 100K questions. To make QuAC, one crowd worker asked a second crowd worker questions regarding a Wikipedia article, and the second crowd worker responded by picking a text span. [7]

NewsQA has 100K questions based on 10K CNN articles. One group of crowd employees created questions based on a headline and synopsis, while another group of workers discovered the solution in the article. [8]

Table 2: Previously reported F1 scores.

Dataset	Reference
SQuAD	76.3
TriviaQA	56.3
NQ	52.7
QuAC	54.4
NewsQA	66.8

In Table 2, previously published BERT results are provided. When we make changes to match SQuAD, we get differences.

Our datasets differ significantly in terms of genre and construction method. SQuAD exceeds all other models in terms of size and accuracy. Because I believe that a strong reading comprehension model should be able to handle question responding regardless of dataset variances, we compare across all five datasets.

4. Methods and materials

Two main works that my project is based on are Unified Language Model (UniLM) and spaCy. SpaCy was used for 'smart' data extraction from the passages for question generation, whereas UniLM was used for actual question generation.

4.1. UniLM

UniLM was used for actual question generation. There are three major benefits to the planned UNILM. To begin with, the unified pre-training approach results in a single Transformer LM that employs shared parameters and architecture for different types of LMs, eliminating the need to train and host many LMs separately. Second, parameter sharing makes learned text representations broader since they are jointly optimized for different language modeling objectives that employ context in different ways, reducing overfitting to a single LM task. Third, UNILM's employment as a sequence-to-sequence LM makes it a suitable candidate for NLG activities like abstractive summarization and question creation, in addition to its applicability to NLU tasks.

Table 3: Extractive QA results on the SQuAD development set.

Model	Exact Match (EM)	F1 score
RMR+ELMo	71.4	73.7
BERTLARGE	78.9	81.8

UNILM	80.5	83.4

The results on SQuAD 2.0 are reported in Table 3, where we compare two models in Exact Match (EM) and F1 score. RMR+ELMo is an LSTM-based question answering model augmented with pre-trained language representation. BERTLARGE is a cased model, fine-tuned on the SQuAD training data for 3 epochs, with batch size 24, and maximum length 384. UNILM is fine-tuned in the same way as BERTLARGE. We see that UNILM outperforms BERTLARGE. [9]

4.2 SpaCy

Spacy is an open-source library that supports advanced Natural Language Processing tools such as tokenization and light deep learning models for tasks such as tagging and Named Entity Recognition (NER). [10] In this work the NER function of the library is used. The list of all possible NER labels of spaCy are shown in Table 4.

Table 4: SpaCy NER labels

PERSON	People, including fictional ones.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including “%”.
MONEY	Monetary values, including units.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	“first”, “second”, etc.
CARDINAL	Numerals that do not fall under another type.

4. Implementation

There are two main parts of this project: Question Generation, Input Preparation.

4.1 Question Generation

The UniLM model for question generation requires an input file in a specific defined format and returns an output file with generated questions. For possible future expansion of an application the model is run as a separate service based on a Flask, micro web framework for python. Its only function is running the UniLM model. It receives the text, saves it in a txt file in a particular format, runs the model and sends back the questions from the output txt file.

4.1.1 Hyperparameters

In order to launch a model, hyperparameters have to be customized or in other words *tuned.* There is a fine-tuned checkpoint decoding used in project:

```
subprocess.call(
            ['python', '/unilm/unilm-v1/src/biunilm/decode_seq2seq.py',
             '--bert_model', 'bert-large-cased',
             '--new_segment_ids', '--mode', 's2s',
             '--input_file', '/data/answers.txt',
             '--model_recover_path', '/data/qg_model.bin',
             '--max_seq_length', '512', '--max_tgt_length', '48',
             '--batch_size', '16', '--beam_size', '1', '--length_penalty', '0',
             '--output_file', '/data/questions2.txt'], env=myenv)
```

Where subprocess is a library in python used to call a decoding itself. Model-recover-path is a path for a Unified Langauge Model which was pre-trained on SQuAD dataset. Other parameters are self-explanatory.

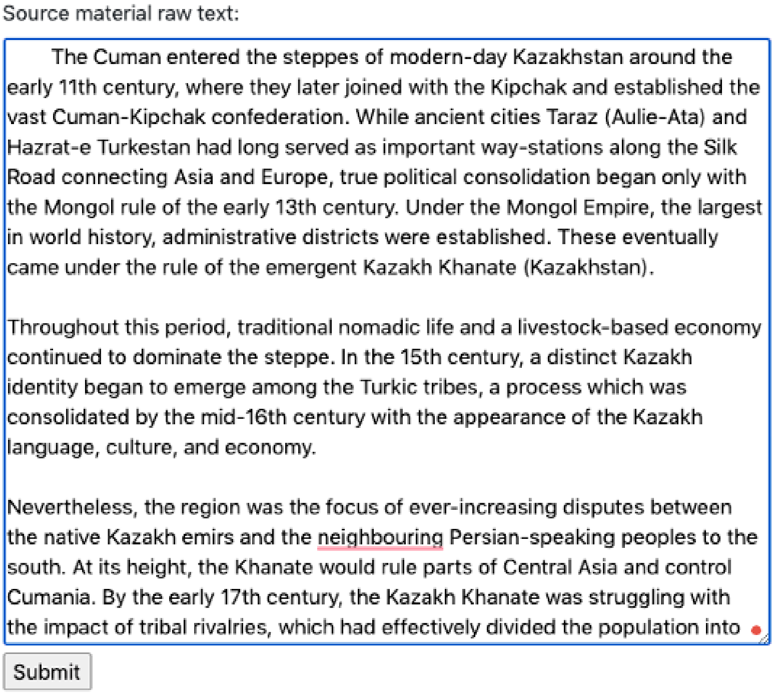
4.2 Input Preparation

UniLM models require input in a format <Context>[SEP]<Answer>, where <context> is the context of the answer and <answer> is the answer for the desired question. If we want to have a question about Genghis Khan’s year of death, we need the data to be in the following format: Genghis Khan died in 1227 after defeating the Western Xia. [SEP] 1227.

Named Entity Recognition (NER) was used. SpaCy allows to get the named entities of the text and having those named entities, shown in Table 4, allow to get more reasonable input for the UniLM model. Firstly, we discard all sentences that don't have those NEs. Then we randomly sample the amount that we need and for each sentence we include a couple of previous sentences as a context to get a broader picture.

7. Results

The final project is an app that runs on Flask server. Users can upload reading material for question generation. From Figure 1 and 2 you can see the screenshots for text entering and test generation, respectively. In this example text material is about Kazakh Khanate, it was taken from Wikipedia page about Kazakhstan [11].



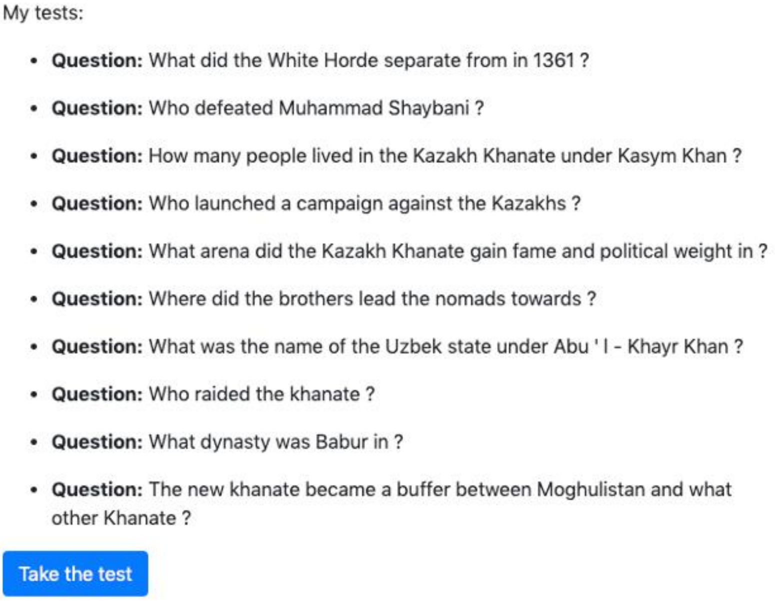
Source material raw text:

The Cuman entered the steppes of modern-day Kazakhstan around the early 11th century, where they later joined with the Kipchak and established the vast Cuman-Kipchak confederation. While ancient cities Taraz (Aulie-Ata) and Hazrat-e Turkestan had long served as important way-stations along the Silk Road connecting Asia and Europe, true political consolidation began only with the Mongol rule of the early 13th century. Under the Mongol Empire, the largest in world history, administrative districts were established. These eventually came under the rule of the emergent Kazakh Khanate (Kazakhstan).

Throughout this period, traditional nomadic life and a livestock-based economy continued to dominate the steppe. In the 15th century, a distinct Kazakh identity began to emerge among the Turkic tribes, a process which was consolidated by the mid-16th century with the appearance of the Kazakh language, culture, and economy.

Nevertheless, the region was the focus of ever-increasing disputes between the native Kazakh emirs and the neighbouring Persian-speaking peoples to the south. At its height, the Khanate would rule parts of Central Asia and control Cumania. By the early 17th century, the Kazakh Khanate was struggling with the impact of tribal rivalries, which had effectively divided the population into

Figure 1: Entering raw text material to process



My tests:

- **Question:** What did the White Horde separate from in 1361 ?
- **Question:** Who defeated Muhammad Shaybani ?
- **Question:** How many people lived in the Kazakh Khanate under Kasym Khan ?
- **Question:** Who launched a campaign against the Kazakhs ?
- **Question:** What arena did the Kazakh Khanate gain fame and political weight in ?
- **Question:** Where did the brothers lead the nomads towards ?
- **Question:** What was the name of the Uzbek state under Abu ' l - Khayr Khan ?
- **Question:** Who raided the khanate ?
- **Question:** What dynasty was Babur in ?
- **Question:** The new khanate became a buffer between Moghulistan and what other Khanate ?

Figure 2: Generated questions from given text material

8. Conclusion

To sum up, a web application that automatically generates and assesses test questions and answers is proposed. First off, several public datasets were compared by amount, accuracy, performance by known models as BERT. It is concluded to use exactly SQuAD dataset, because it has outperformed other datasets. Thus, the system I propose uses SQuAD dataset, state-of-the-art Natural Language Processing (NLP) technologies: Unified Language Model (UniLM) and spaCy. SpaCy was used for 'smart' data extraction from the passages for question generation, while UniLM was used for actual question generation. Proposed solution will work on generating open questions.

In future, work will be dedicated to several details. Firstly, extractive summarization models could be used to prioritize important questions first to be generated. Secondly, to solve problems with similarity scores for numbers. Finally, improve the user interface of an application.

References

1 Daniel W. Otter, Julian R. Medina, Jugal K. Kalita. *A Survey of the Usages of Deep Learning for Natural Language Processing*, 2019. pp. 1-2.

2 D. Khurana , A. Koli, K. Khatter, S. Sukhdev. *Natural Language Processing: State of The Art, Current Trends and Challenges. Department of Computer Science and Engineering Manav Rachna International University*, 2017. pp. 1-25

3 M. Zock, G. Adorni. *Trends in Natural Language Generation: an Artificial Intelligence Perspective*, 1996. pp. 1-16

4 Pranav Rajpurkar, Robin Jia, Percy Liang. *Know what you don't know: Unanswerable questions for SQuAD*, 2018. pp. 784–789

5 Mandar Joshi, Eunsol Choi, Daniel Weld, Luke Zettlemoyer. *TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension*, 2017. pp. 1601–1611

6 Tom Kwiatkowski, Jennimaria Palomaki, Illia Polosukhin, Olivia Redfield, Jacob Devlin, Michael Collins, Ankur Parikh, Chris Alberti, Andrew M. Dai, Danielle Epstein, Kenton Lee, Kristina Toutanova, Matthew Kelcey, Ming-Wei Chang, Jakob Uszkoreit, Llion Jones, Quoc Le, Slav Petrov. *Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics*, 2019. pp. 453–466.

7 Eunsol Choi, Mohit Iyyer, He He, Mark Yatskar, Yejin Choi, Wentau Yih, Percy Liang, Luke Zettlemoyer. *QuAC: Question answering in context*, 2018. pp. 2174–2184.

8 Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, Kaheer Suleman. *NewsQA: A machine comprehension dataset*, 2017. pp. 191–200

9 Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, Hsiao-Wuen Hon. *Unified Language Model*

Pre-training for Natural Language Understanding and Generation, 2019. pp. 1-14

10 SpaCy. *Industrial-strength Natural Language Processing in Python*, 2020. URL: https://spacy.io/

11 Wikipedia. *Kazakhstan*, 2020. URL: https://en.wikipedia.org/wiki/Kazakhstan