Beyond English-Only Reading Comprehension: Experiments in Zero-Shot Multilingual Transfer for Bulgarian

Momchil Hardalov[1]   **Ivan Koychev**[1]   **Preslav Nakov**[2]
[1]Sofia University "St. Kliment Ohridski", Bulgaria,
[2]Qatar Computing Research Institute, HBKU, Qatar,
`{hardalov, koychev}@fmi.uni-sofia.bg`
`pnakov@hbku.edu.qa`

Abstract

Recently, reading comprehension models achieved near-human performance on large-scale datasets such as SQuAD, CoQA, MS Macro, RACE, etc. This is largely due to the release of pre-trained contextualized representations such as BERT and ELMo, which can be fine-tuned for the target task. Despite those advances and the creation of more challenging datasets, most of the work is still done for English. Here, we study the effectiveness of multilingual BERT fine-tuned on large-scale English datasets for reading comprehension (e.g., for RACE), and we apply it to Bulgarian multiple-choice reading comprehension. We propose a new dataset containing 2,221 questions from matriculation exams for twelfth grade in various subjects —history, biology, geography and philosophy—, and 412 additional questions from online quizzes in history. While the quiz authors gave no relevant context, we incorporate knowledge from Wikipedia, retrieving documents matching the combination of question + each answer option. Moreover, we experiment with different indexing and pre-training strategies. The evaluation results show accuracy of 42.23%, which is well above the baseline of 24.89%.

1 Introduction

The ability to answer questions is natural to humans, independently of their native language, and, once learned, it can be easily transferred to another language. After understanding the question, we typically depend on our background knowledge, and on relevant information from external sources.

Machines do not have the reasoning ability of humans, but they are still able to learn concepts. The growing interest in teaching machines to answer questions posed in natural language has led to the introduction of various new datasets for different tasks such as reading comprehension, both extractive, e.g., span-based (Nguyen et al., 2016; Trischler et al., 2017; Joshi et al., 2017; Rajpurkar et al., 2018; Reddy et al., 2019), and non-extractive, e.g., multiple-choice questions (Richardson et al., 2013; Lai et al., 2017; Clark et al., 2018; Mihaylov et al., 2018; Sun et al., 2019a). Recent advances in neural network architectures, especially the raise of the Transformer (Vaswani et al., 2017), and better contextualization of language models (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2018; Grave et al., 2018; Howard and Ruder, 2018; Radford et al., 2019; Yang et al., 2019b; Dai et al., 2019) offered new opportunities to advance the field.

Here, we investigate skill transfer from a high-resource language, i.e., English, to a low-resource one, i.e., Bulgarian, for the task of multiple-choice reading comprehension. Most previous work (Pan et al., 2018; Radford et al., 2018; Tay et al., 2018; Sun et al., 2019b) was monolingual, and a relevant context for each question was available a priori. We take the task a step further by exploring the capability of a neural comprehension model in a multilingual setting using external commonsense knowledge. Our approach is based on the multilingual cased BERT (Devlin et al., 2019) fine-tuned on the RACE dataset (Lai et al., 2017), which contains over 87,000 English multiple-choice school-level science questions. For evaluation, we build a novel dataset for Bulgarian. We further experiment with pre-training the model over stratified Slavic corpora in Bulgarian, Czech, and Polish Wikipedia articles, and Russian news, as well as with various document retrieval strategies.

Finally, we address the resource scarceness in low-resource languages and the absence of question contexts in our dataset by extracting relevant passages from Wikipedia articles.

Our contributions are as follows:

- We introduce a new dataset for reading comprehension in a low-resource language such as Bulgarian. The dataset contains a total of 2,636 multiple-choice questions without contexts from matriculation exams and online quizzes. These questions cover a large variety of science topics in biology, philosophy, geography, and history.

- We study the effectiveness of zero-shot transfer from English to Bulgarian for the task of multiple-choice reading comprehension, using Multilingual and Slavic BERT (Devlin et al., 2019), fine-tuned on large corpora, such as RACE (Lai et al., 2017).

- We design a general-purpose pipeline[1] for extracting relevant contexts from an external corpus of unstructured documents using information retrieval.

The rest of this paper is organized as follows: The next section presents related work. Section 3 describes our approach. Details about the newly-proposed multiple-choice Bulgarian dataset are given in Section 4. All experiments are described in Section 5. Finally, Section 6 concludes and points to possible directions for future work.

2 Related Work

2.1 Machine Reading Comprehension

The growing interest in machine reading comprehension (MRC) has led to the release of various datasets for both extractive (Nguyen et al., 2016; Trischler et al., 2017; Joshi et al., 2017; Rajpurkar et al., 2018; Reddy et al., 2019) and non-extractive (Richardson et al., 2013; Peñas et al., 2014; Lai et al., 2017; Clark et al., 2018; Mihaylov et al., 2018; Sun et al., 2019a) comprehension. Our work primarily focuses on the non-extractive multiple-choice type, designed by educational experts, since their task is very close to our newly-proposed dataset, and are expected to be well-structured and error-free (Sun et al., 2019a).

These datasets brought a variety of models and approaches. The usage of external knowledge has been an interesting topic, e.g., Chen et al. (2017a) used Wikipedia knowledge for answering open-domain questions, Pan et al. (2018) applied entity discovery and linking as a source of prior knowledge. Sun et al. (2019b) explored different reading strategies such as back and forth reading, highlighting, and self-assessment. Ni et al. (2019) focused on finding essential terms and removing distraction words, followed by reformulation of the question, in order to find better evidence before sending a query to the MRC system. A simpler approach was presented by Clark et al. (2016), who leveraged information retrieval, corpus statistics, and simple inference over a semi-automatically constructed knowledge base for answering fourth-grade science questions.

Current state-of-the-art approaches in machine reading comprehension are grounded on transfer learning and fine-tuning of language models (Peters et al., 2018; Conneau et al., 2018; Devlin et al., 2019). Yang et al. (2019a) presented an open-domain extractive reader based on BERT (Devlin et al., 2019). Radford et al. (2018) used generative pre-training of a Transformer (Vaswani et al., 2017) as a language model, transferring it to downstream tasks such as natural language understanding, reading comprehension, etc.

Finally, there has been a Bulgarian MRC dataset (Peñas et al., 2012). It was used by Simov et al. (2012), who converted the question-answer pairs to declarative sentences, and measured their similarity to the context, transforming both to a bag of linguistic units: lemmata, POS tags, and dependency relations.

2.2 (Zero-Shot) Multilingual Models

Multilingual embeddings helped researchers to achieve new state-of-the-art results on many NLP tasks. While many pre-trained model (Grave et al., 2018; Devlin et al., 2019; Lample and Conneau, 2019) are available, the need for task-specific data in the target language still remains. Learning such models is language-independent, and representations for common words remain close in the latent vector space for a single language, albeit unrelated for different languages. A possible approach to overcome this effect is to learn an alignment function between spaces (Artetxe and Schwenk, 2018; Joty et al., 2017).

---

[1]The dataset and the source code are available at `http://github.com/mhardalov/bg-reason-BERT`

448

Moreover, zero-shot application of fine-tuned multilingual language models (Devlin et al., 2019; Lample and Conneau, 2019) on XNLI (Conneau et al., 2018), a corpus containing sentence pairs annotated with textual entailment and translated into 14 languages, has shown very close results to such by a language-specific model.

Zero-shot transfer and multilingual models had been a hot topic in (neural) machine translation (MT) in the past several years. Johnson et al. (2017) introduced a simple tweak to a standard sequence-to-sequence (Sutskever et al., 2014) model by adding a special token to the encoder's input, denoting the target language, allowing a zero-shot learning for new language pairs. Recent work in zero-resource translation outlined different strategies for learning to translate without having a parallel corpus between the two target languages. First, a many-to-one approach was adopted by Firat et al. (2016) based on building a corpus from a single language paired with many others, allowing simultaneous training of multiple models, with a shared attention layer. A many-to-many relationship between languages was later used by Aharoni et al. (2019), in an attempt to train a single Transformer (Vaswani et al., 2017) model.

Pivot-language approaches can also be used to overcome the lack of parallel corpora for the source–target language pair. Chen et al. (2017b) used a student-teacher framework to train an NMT model, using a third language as a pivot. A similar idea was applied to MRC by Asai et al. (2018), who translated each question to a pivot language, and then found the correct answer in the target language using soft-alignment attention scores.

3 Model

Our model has three components: (*i*) a context retrieval module, which tries to find good explanatory passages for each question-answer pair, from a corpus of non-English documents, as described in Section 3.1, (*ii*) a multiple-choice reading comprehension module pre-trained on English data and then applied to the target language in a zero-shot fashion, i.e., without further training or additional fine-tuning, to a target (non-English) language, as described in Section 3.2, and (*iii*) a voting mechanism, described in Section 3.3, which combines multiple passages from (*i*) and their scores from (*ii*) in order to obtain a single (most probable) answer for the target question.

3.1 Context Retriever

Most public datasets for reading comprehension (Richardson et al., 2013; Lai et al., 2017; Sun et al., 2019a; Rajpurkar et al., 2018; Reddy et al., 2019; Mihaylov et al., 2018) contain not only questions with possible answers, but also an evidence passage for each question. This limits the task to question answering over a piece of text, while an open-domain scenario is much more challenging and much more realistic. Moreover, a context in which the answer can be found is not easy to retrieve, sometimes even for a domain expert. Finally, data scarceness in low-resource languages poses further challenges for finding resources and annotators.

In order to enable search for appropriate passages for non-English questions, we created an inverted index from Wikipedia articles using Elasticsearch.[2] We used the original dumps for the entire Wikipage,[3] and we preprocessed the data leaving only plain textual content, e.g., removing links, HTML tags, tables, etc. Moreover, we split the article's body using two strategies: a sliding window and a paragraph-based approach. Each text piece with its corresponding article title was processed by applying word-based tokenization, lowercasing, stop-words removal, stemming (Nakov, 2003; Savoy, 2007), and n-gram extraction. Finally, the matching between a question and a passage was done using cosine similarity and BM25 (Robertson and Zaragoza, 2009).

3.2 BERT for Multiple-Choice RC

The recently-proposed BERT (Devlin et al., 2019) framework is applicable to a vast number of NLP tasks. A shared characteristic between all of them is the form of the input sequences: a single sentence or a pair of sentences separated by the [SEP] special token, and a classification token ([CLS]) added at the beginning of each example. In contrast, the input for multiple-choice reading comprehension questions is assembled by three sentence pieces, i.e., context passage, question, and possible answer(s). Our model follows a simple strategy of concatenating the option (candidate answer) at the end of a question. Following the notation of Devlin et al. (2019), the input sequence can be written as follows:

[CLS] Passage [SEP] Question + Option [SEP]

---

[2] http://www.elastic.co/
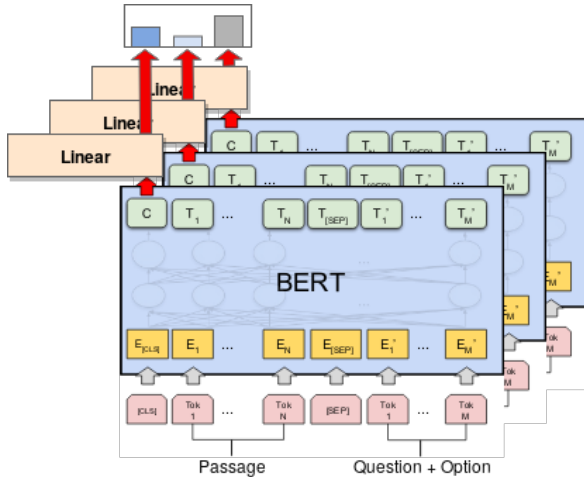[3] http://dumps.wikimedia.org/

Figure 1: BERT for multiple-choice reasoning.

As recommended by Devlin et al. (2019), we introduce a new task-specific parameter vector L, $L \in \mathbb{R}^H$, where H is the hidden size of the model. In order to obtain a score for each passage-question-answer triplet, we take the dot product between L and the final hidden vector for the classification token ([CLS]), thus ending up with N unbounded numbers: one for each option. Finally, we normalize the scores by adding a softmax layer, as shown in Figure 1. During fine-tuning, we optimize the model's parameters by maximizing the log-probability of the correct answer.

3.3 Answer Selection Strategies

Finding evidence passages that contain information about the correct answer is crucial for reading comprehension systems. The context retriever may be extremely sensitive to the formulation of a question. The latter can be very general, or can contain insignificant rare words, which can bias the search. Thus, instead of using only the first-hit document, we should also evaluate lower-ranked ones. Moreover, knowing the answer candidates can enrich the search query, resulting in improved, more answer-oriented passages. This approach leaves us with a set of contexts that need to be evaluated by the MRC model in order to choose a single correct answer. Prior work suggests several different strategies: Chen et al. (2017a) used the raw predicted probability from a recurrent neural network (RNN), Yang et al. (2019a) tuned a hyper-parameter to balance between the retriever score and the reading model's output, while Pan et al. (2018) and Ni et al. (2019) concatenated the results from sentence-based retrieval into a single contextual passage.

In our experiments below, we adopt a simple summing strategy. We evaluate each result from the context retriever against the question and the possible options (see Section 3.2 for more details), thus obtaining a list of raw probabilities. We found empirically that explanatory contexts assign higher probability to the related answer, while general or uninformative passages lead to stratification of the probability distribution over the answer options. We formulate this as follows:

$$Pr(a_j|p;q) = \frac{exp(BERT(p, q + a_j))}{\sum_{j'} exp(BERT(p, q + a_{j'}))}, \quad (1)$$

where p is a passage, q is a question, A is the set of answer candidates, and $a_j \in A$.

We select the final answer as follows:

$$Ans = \arg\max_{a \in A} \sum_{p \in P} Pr(A|p;q) \quad (2)$$

4 Data

Our goal is to build a task for a low-resource language, such as Bulgarian, as close as possible to the multiple-choice reading comprehension setup for high-resource languages such as English. This will allow us to evaluate the limitations of transfer learning in a multilingual setting. One of the largest datasets for this task is RACE (Lai et al., 2017), with a total of 87,866 training questions with four answer candidates for each. Moreover, there are 25,137 contexts mapped to the questions and their correct answers.

While there exist many datasets for reading comprehension, most of them are in English, and there are a very limited number in other languages (Peñas et al., 2012, 2014). Hereby, we collect our own dataset for Bulgarian, resulting in 2,633 multiple-choice questions, without contexts, from different subjects: biology (16.6%), philosophy (23.93%), geography (23.24%), and history (36.23%). Table 2 shows an example question with candidate answers chosen to represent best each category. We use green to mark the correct answer, and bold for the question category. For convenience all the examples are translated to English.

Table 1 shows the distribution of questions per subject category, the length (in words) for both the questions and the options (candidate answers), and the vocabulary richness, measured in terms of unique words. The first part of the table presents statistics about our dataset, while the second part is a comparison to RACE (Lai et al., 2017).

Domain	#QA-pairs	#Choices	Len Question	Len Options	Vocabulary Size
12th Grade Matriculation Exam					
Biology	437	4	10.4	2.6	2, 414 (12, 922)
Philosophy	630	4	8.9	2.9	3, 636 (20, 392)
Geography	612	4	12.8	2.5	3, 239 (17, 668)
History	542	4	23.7	3.6	5, 466 (20, 456)
Online History Quizzes					
Bulgarian History	229	4	14.0	2.8	2, 287 (10, 620)
PzHistory	183	3	38.9	2.4	1, 261 (7, 518)
Overall	2, 633	3.9	15.7	2.9	13, 329 (56, 104)
RACE Train - Mid and High School					
RACE-M	25, 421	4	9.0	3.9	32, 811
RACE-H	62, 445	4	10.4	5.8	125, 120
Overall	87, 866	4	10.0	5.3	136, 629

Table 1: Statistics about our Bulgarian dataset compared to the RACE dataset.

(**Biology**) The thick coat of mammals in winter is an example of:
A. physiological adaptation
B. behavioral adaptation
C. genetic adaptation
D. morphological adaptation

(**Philosophy**) According to relativism in ethics:
A. there is only one moral law that is valid for all
B. there is no absolute good and evil
C. people are evil by nature
D. there is only good, and the evil is seeming

(**Geography**) Which of the assertions about the economic specialization of the Southwest region is true?
A. The ratio between industrial and agricultural production is 15:75
B. Lakes of glacial origin in Rila and Pirin are a resource for the development of tourism
C. Agricultural specialization is related to the cultivation of grain and ethereal-oil crops
D. The rail transport is of major importance for intra-regional connections

(**History**) Point out the concept that is missed in the text of the Turnovo Constitution: „Art. 54 All born in Bulgaria, also those born elsewhere by parents Bulgarian _____, count as _____ of the Bulgarian Principality. Art. 78 Initial teaching is free and obligatory for all _____ of the Bulgarian Principality."
A. residents
B. citizents
C. electors
D. voters

(**History Quiz**) Sofroniy Vrachanski started a family that plays a big role in the history of the Bulgarian National Revival. What is its name?
A. Georgievi
B. Tapchileshtovi
C. Bogoridi
D. Palauzovi

Table 2: Example questions, one per subject, from our Bulgarian dataset. The correct answer is marked in green.

We divided the Bulgarian questions into two groups based on the question's source. The first group (*12th Grade Matriculation Exam*) was collected from twelfth grade matriculation exams created by the Ministry of Education of Bulgaria in the period 2008–2019. Each exam contains thirty multiple-choice questions with four possible answers per question. The second set of questions (*Online History Quizzes*) are history-related and are collected from online quizzes. While they are not created by educators, the questions are still challenging and well formulated. Furthermore, we manually filtered out questions with non-textual content (i.e., pictures, paintings, drawings, etc.), ordering questions (i.e., order the historical events), and questions involving calculations (i.e., how much X we need to add to Y to arrive at Z).

Table 1 shows that history questions in general contain more words (14.0–38.9 on average), compared to other subjects (8.9–12.8 on average). A tangible difference in length compared to other subjects is seen for *12th grade History* and *PzHistory*, due to the large number of quotes, and document pieces contained in questions from these two groups. Also, the average question length is 15.7, which is longer compared to the RACE dataset with 10.0. On the other hand, the option lengths per subject category in our dataset follow a narrower distribution. They fall in the interval between 2.5 and 2.9 words on average, expect for *12th grade History*, with 3.6 words. Here, we note a significant difference compared to the option lengths in RACE, which tend to be 2.4 words longer on average – 5.3 for RACE vs. 2.9 for ours.

Finally, we examine the vocabulary richness of the two datasets. The total number of unique words is shown in the last column of Table 1 (Vocab Size). For our dataset, there are two numbers per row: the first one shows statistics based on the question–answer pairs only, while the second one, enclosed in parentheses, measures the vocabulary size including the extracted passages by the Context Retriever. The latter number is a magnitude estimate rather then a concrete number, since its upper limit is the number of words in Wikipedia, and it can vary for different retrieval strategies.

5 Experiments and Evaluation

5.1 BERT Fine-Tuning

We divide the fine-tuning into two groups of models (*i*) Multilingual BERT, and (*ii*) Slavic BERT. Table 3 below presents the results in the multiple-choice comprehension task on the dev dataset from RACE (Lai et al., 2017).

#Epoch	RACE-M	RACE-H	Overall
BERT 1	64.21	53.66	56.73
BERT 2	68.80	57.58	60.84
BERT 3	69.15	58.43	61.55
Slavic 2	53.55	44.48	47.12
Slavic 3	57.38	46.88	49.94

Table 3: Accuracy measured on the dev RACE dataset after each training epoch.

Multilingual BERT As our initial model, we use $BERT_{base}$, Multilingual Cased which is pre-trained on 104 languages, and has 12-layers, 768-hidden units per layer, 12-heads, and a total of 110M parameters. We further fine-tune the model on RACE (Lai et al., 2017) for 3 epochs saving a checkpoint after each epoch. We use a batch size of 8, a max sequence size of 320, and a learning rate of 1e-5.

Slavic BERT The Slavic model[4] was built using transfer learning from the Multilingual BERT model to four Slavic languages: Bulgarian, Czech, Polish, and Russian. In particular, the Multilingual BERT model was fine-tuned on a stratified dataset of Russian news and Wikipedia articles for the other languages. We use this pre-trained Slavic BERT model, and we apply the same learning procedure as for *Multilingual BERT*.

---

[4]http://github.com/deepmipt/
Slavic-BERT-NER

5.2 Wikipedia Retrieval and Indexing

Here, we discuss the retrieval setup (see Section 3.1 for more details). We use the Bulgarian dump of Wikipedia from 2019-04-20, with a total of 251,507 articles. We index each article title and body in plain text, which we call a *passage*. We further apply additional processing for each field:

- *ngram*: word-based 1–3 grams;
- *bg*: lowercased, stop-words removed (from Lucene), and stemmed (Savoy, 2007);
- *none*: bag-of-words index.

We ended up using a subset of four fields from all the possible analyzer-field combinations, namely *title.bg, passage, passage.bg,* and *passage.ngram*. We applied Bulgarian analysis on the *title* field only as it tends to be short and descriptive, and thus very sensitive to noise from stop-words, which is in contrast to questions that are formed mostly of stop-words, e.g., *what*, *where*, *when*, *how*.

For indexing the Wikipedia articles, we adopt two strategies: sliding window and paragraph. In the window-based strategy, we define two types of splits: small, containing 80-100 words, and large, of around 300 words. In order to obtain indexing chunks, we define a window of size K, and a stride equal to one forth of K. Hence, each $\frac{K}{4}$ characters, which is the size of the stride, are contained into four different documents. The paragraph-based strategy divides the article by splitting it using one or more successive newline characters ([\n]+) as a delimiter. We avoid indexing entire documents due to their extensive length, which can be far beyond the maximum length that BERT can take as an input, i.e., 320 word pieces (see Section 5.1 for the more details). Note that extra steps are needed in order to extract a proper passage from the text. Moreover, the amount of facts in the Wikipedia articles that are unrelated to our questions give rise to false positives since the question is short and term-unspecific.

Finally, we use a list of top-N hits for each candidate answer. Thus, we have to execute an additional query for each question + option combination, which may result in duplicated passages, thus introducing an implicit bias towards the candidates they support. In order to mitigate this effect, during the answer selection phase (see Section 3.3), we remove all duplicate entries, keeping a single instance.

Setting	Accuracy
Random	24.89
Train for 3 epochs	–
+ window & title.bg & pass.ngram	29.62
+ passage.bg & passage	39.35
– title.bg	39.69
+ passage.bg^2	40.26
+ title.bg^2	40.30
+ bigger window	36.54
+ paragraph split	42.23
+ Slavic pre-training	33.27
Train for 1 epoch best	40.26
Train for 2 epochs best	41.89

Table 4: Accuracy on the Bulgarian testset: ablation study when sequentially adding/removing different model components.

5.3 Experimental Results

Here, we discuss the accuracy of each model on the original English MRC task, followed by experiments in zero-shot transfer to Bulgarian.

English Pre-training for MCRC. Table 3 presents the change in accuracy on the original English comprehension task, depending on the number of training epochs. In the table, "BERT" refers to the Multilingual BERT model, while "Slavic" stands for BERT with Slavic pre-training. We further fine-tune the models on the RACE dataset. Next, we report their performance in terms of accuracy, following the notation from (Lai et al., 2017). Note that the questions in RACE-H are more complex than those in RACE-M. The latter has more word matching questions and fewer reasoning questions. The final column in the table, *Overall*, shows the accuracy calculated over all questions in the RACE testset. We train both setups for three epochs and we report their performance after each epoch. We can see a positive correlation between the number of epochs and the model's accuracy. We further see that the Slavic BERT performs far worse on both RACE-M and RACE-H, which suggests that the change of weights of the model towards Slavic languages has led to catastrophic forgetting of the learned English syntax and semantics. Thus, it should be expected that the adaptation to Slavic languages would yield decrease in performance for English. What matters though is whether this helps when testing on Bulgarian, which we explore next.

Zero-Shot Transfer. Here, we assess the performance of our model when applied to Bulgarian multiple-choice reading comprehension. Table 4 presents an ablation study for various components. Each line denotes the type of the model, and the addition (+) or the removal (–) of a characteristic from the setup in the previous line. The first line shows the performance of a baseline model that chooses an option uniformly at random from the list of candidate answers for the target question. The following rows show the results for experiments conducted with a model trained for three epochs on RACE (Lai et al., 2017).

Our basic model uses the following setup: Wikipedia pages indexed using a small sliding window (400 characters, and stride of 100 characters), and context retrieval over two fields: Bulgarian analyzed title (*text.bg*), and word n-grams over the passage (*passage.ngram*). This setup yields 29.62% accuracy, and it improves over the random baseline by 4.73% absolute. We can think of it as a non-random baseline for further experiments. Next, we add two more fields to the IR query: passage represented as a bag of words (named *passage*), and Bulgarian analyzed (*passage.bg*), which improves the accuracy by additional 10%, arriving at 39.35%. The following experiment shows that removing the *title.bg* field does not change the overall accuracy, which makes it an insignificant field for searching. Further, we add double weight on *passage.bg*, (shown as ^2), which yields 1% absolute improvement.

From the experiments described above, we found the best combination of query fields to be *title.bulgarian^2, passage.ngram, passage, passage.bulgarian^2*, where the *title* has a minor contribution, and can be sacrificed for ease of computations and storage. Fixing the best query fields, allowed us to evaluate other indexing strategies, i.e., bigger window (size 1,600, stride 400) with accuracy 36.54%, and paragraph splitting, with which we achieved our highest accuracy of 42.23%. This is an improvement of almost 2.0% absolute over the small sliding window, and 5.7% over the large one.

Next, we examined the impact of the Slavic BERT. Surprisingly, it yielded 9% absolute drop in accuracy compared to the multi-lingual BERT. This suggests that the latter already has enough knowledge about Bulgarian, and thus it does not need further adaptation to Slavic languages.
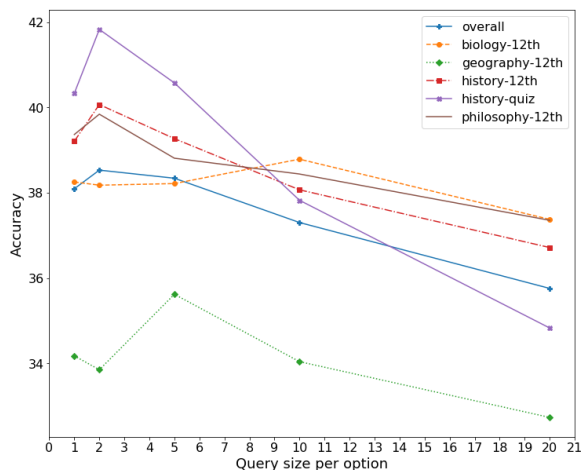
Figure 2: Accuracy per question category based on the number of query results per answer option.

Next, we study the impact of the number of fine-tuning epochs on the model's performance. We observe an increase in accuracy as the number of epochs grows, which is in line with previously reported results for English tasks. While this correlation is not as strong as for the original RACE task (see Table 3 for comparison), we still observe 1.6% and 0.34% absolute increase in accuracy for epochs 2 and 3, respectively, compared to epoch 1. Note that we do not go beyond three epochs, as previous work has suggested that 2-3 fine-tuning epochs are enough (Devlin et al., 2019), and after that, there is a risk of catastrophic forgetting of what was learned at pre-training time (note that we have already seen such forgetting with the Slavic BERT above).

We further study the impact of the size of the results list returned by the retriever on the accuracy for the different categories. Figure 2 shows the average accuracy for a given query size S_q over all performed experiments, where $S_q \in \{1, 2, 5, 10, 20\}$. We can see in Figure 2 that longer query result lists (i.e., containing more than 10 results) per answer option worsen the accuracy for all categories, except for *biology*, where we see a small peak at length 10, while still the best overall results for this category is achieved for a result list of length 5. A single well-formed maximum at length 2 is visible for *history* and *philosophy*. With these two categories being the biggest ones, the cap at the same number of queries for the overall accuracy is not a surprise. The per-category results for the experiments are discussed in more detail in Appendix A.

We can see that the highest accuracy is observed for *history*, particularly for online quizzes, which are not designed by educators and are more of a word-matching nature rather then a reasoning one (see Table 2). Finally, *geography* appears to be the hardest category with only 38.73% accuracy: 3.5% absolute difference compared to the second-worst category. The performance for this subject is also affected differently by changes in query result length: the peak is at lengths 5 and 10, while there is a drop for length 2. A further study of the model's behavior can be found in Appendix B.

6 Conclusion and Future Work

We studied the task of multiple-choice reading comprehension for low-resource languages, using a newly collected Bulgarian corpus with 2,633 questions from matriculation exams for twelfth grade in history and biology, and online exams in history without explanatory contexts. In particular, we designed an end-to-end approach, on top of a multilingual BERT model (Devlin et al., 2019), which we fine-tuned on large-scale English reading comprehension corpora, and open-domain commonsense knowledge sources (Wikipedia). Our main experiments evaluated the model when applied to Bulgarian in a zero-shot fashion. The experimental results found additional pre-training on the English RACE corpus to be very helpful, while pre-training on Slavic languages to be harmful, possibly due to catastrophic forgetting. Paragraph splitting, n-grams, stop-word removal, and stemming further helped the context retriever to find better evidence passages, and the overall model to achieve accuracy of up to 42.23%, which is well above the baselines of 24.89% and 29.62%.

In future work, we plan to make use of reading strategies (Sun et al., 2019b), linked entities (Pan et al., 2018), concatenation and reformulation of passages and questions (Simov et al., 2012; Clark et al., 2016; Ni et al., 2019), as well as re-ranking of documents (Nogueira and Cho, 2019).

Acknowledgments

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the Conference of the North American Chapter of ACL*. Minneapolis, MN, USA, NAACL-HLT '19, pages 3874–3884.

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464* .

Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *arXiv preprint arXiv:1809.03275* .

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017a. Reading Wikipedia to answer open-domain questions. In *Proceedings of the Meeting of the Association for Computational Linguistics*. Vancouver, Canada, ACL '17, pages 1870–1879.

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017b. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada, ACL '17, pages 1925–1935.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457* .

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *Proceedings of the 13th AAAI Conference on Artificial Intelligence*. Phoenix, AZ, USA, AAAI '16, pages 2580–2586.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, EMNLP '18, pages 2475–2485.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the Meeting of the Association for Computational Linguistics*. Florence, Italy, ACL '19, pages 2978–2988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Minneapolis, MN, USA, NAACL-HLT '19, pages 4171–4186.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*. Austin, TX, USA, EMNLP '16, pages 268–277.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Conference on Language Resources and Evaluation*. Miyazaki, Japan, LREC '18.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Melbourne, Australia, ACL '18, pages 328–339.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics* 5:339–351.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada, ACL '17, pages 1601–1611.

Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. 2017. Cross-language learning with adversarial neural networks. In *Proc. of the Conference on Computational Natural Language Learning*. Vancouver, Canada, CoNLL '17, pages 226–237.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark, EMNLP '17, pages 785–794.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291* .

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, EMNLP '18, pages 2381–2391.

Preslav Nakov. 2003. Building an inflectional stemmer for Bulgarian. In *Proceedings of the 4th International Conference Conference on Computer Systems and Technologies: E-Learning*. Rousse, Bulgaria, CompSysTech '03, pages 419–424.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*. Barcelona, Spain, CoCo@NIPS '16.

Jianmo Ni, Chenguang Zhu, Weizhu Chen, and Julian McAuley. 2019. Learning to attend on essential terms: An enhanced retriever-reader model for open-domain question answering. In *Proceedings of the Conference of the North American Chapter of ACL*. Minneapolis, MN, USA, NAACL-HLT '19, pages 335–344.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085* .

Xiaoman Pan, Kai Sun, Dian Yu, Heng Ji, and Dong Yu. 2018. Improving question answering with external knowledge. *arXiv preprint:1902.00993* .

Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Richard Sutcliffe, Corina Forascu, Yassine Benajiba, and Petya Osenova. 2012. Overview of QA4MRE at CLEF 2012: Question answering for machine reading evaluation. In *CLEF Working Note Papers*. Rome, Italy, pages 1–24.

Anselmo Peñas, Christina Unger, and Axel-Cyrille Ngonga Ngomo. 2014. Overview of CLEF question answering track 2014. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*. pages 300–306.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*. New Orleans, LA, USA, NAACL-HLT '18, pages 2227–2237.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training .

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners .

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the Meeting of the Association for Computational Linguistics*. Melbourne, Australia, ACL '18, pages 784–789.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics* 7:249–266.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, WA, USA, EMNLP '13, pages 193–203.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.* 3(4):333–389.

Jacques Savoy. 2007. Searching strategies for the Bulgarian language. *Inform. Retrieval* 10(6):509–529.

Kiril Ivanov Simov, Petya Osenova, Georgi Georgiev, Valentin Zhikov, and Laura Tolosi. 2012. Bulgarian question answering for machine reading. In *CLEF Working Note Papers*. Rome, Italy.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019a. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics* 7:217–231.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019b. Improving machine reading comprehension with general reading strategies. In *Proceedings of the North American Chapter of ACL*. Minneapolis, MN, USA, NAACL-HLT '19, pages 2633–2643.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*. Montreal, Canada, NIPS '14, pages 3104–3112.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Multi-range reasoning for machine comprehension. *arXiv preprint arXiv:1803.09074* .

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Vancouver, Canada, RepL4NLP '19, pages 191–200.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Annual Conference on Neural Information Processing Systems*. Long Beach, CA, USA, NIPS '17, pages 5998–6008.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019a. End-to-end open-domain question answering with BERTserini. In *Proceedings of the Conference of the North American Chapter of ACL*. Minneapolis, MN, USA, NAACL-HLT '19, pages 72–77.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019b. XLNet: Generalized autoregressive pre-training for language understanding. *arXiv preprint arXiv:1906.08237* .

Appendix

A Per-Category Results

Table 5 gives an overview, including per-category breakdown, of our parameter tuning experiments. We present the results for some interesting experiments rather then for a full grid search. The first row shows a random baseline for each category. In the following rows, we compare different types of indexing: first, we show the results for a small sliding window (400-character window, and 100-character stride), followed by a big window (1,600-character window, and 400-character stride), and finally for paragraph indexing. We use the same notation as in Section 5. The last group in the table (*Paragraph*) shows the best-performing model, where we mark in bold the highest accuracy for each category. For completeness, we also show the accuracy when using the *Slavic BERT* model for prediction, which yields a 10% drop on average compared to using the *Multilingual BERT*, for each of the categories.

B Case Study

In Table 6, we present the retrieved evidence passages for the example questions in Table 2: we omit the answers, and we only show the questions and the contexts. Each example is separated by a double horizontal line, where the first row is the question starting with "*Q:*", and the following rows contain passages returned by the retriever. For each context, we normalize the raw scores from the comprehension model using Eq. 1 to obtain a probability distribution. We then select an answer using $\arg\max$, according to Eq. 2. In the table, we indicate the correctness of each predicted answer using one of the following symbols before the question:

✓ The question is answered correctly.

✗ An incorrect answer has the highest score.

? Two or more answers have the highest score.

We show the top retrieved result in order to illustrate the model scores over different evidence passages and the quality of the articles. The queries are formed by concatenating the question with an answer option, even though this can lead to duplicate results since some answers can be quite similar or the question's terms could dominate the similarity score.

The questions in Table 6 are from five different categories: biology, philosophy, geography, history, and online quizzes. Each of them has its own specifics and gives us an opportunity to illustrate a different model behavior.

The first question is from the biology domain, and we can see that the text is very general, and so is the retrieved context. The latter talks about *hair* rather than *coat*, and the correct answer (D) *morphological adaptation* is not present in the retrieved text. On the other hand, all the terms are only connected to it, and hence the model assigns high probability to this answer option.

For the second question, from the philosophy domain, there are two related contexts found. The first one is quite short, noisy, and it does not give much information in general. The second paragraph manages to extract the definition of *relativism* and to give good supporting evidence for the correct answer, namely that *there is no absolute good and evil* (B). As a result, this option is assigned high probability. Nevertheless, the incorrect answer *here is only one moral law that is valid for all* (A) is assigned an even higher probability and it wins the voting.

In the third example, from the domain of geography, we see a large number of possible contexts, due to the long and descriptive answers. We can make two key observations: (*i*) the query is drawn in very different directions by the answers, and (*ii*) there is no context for *Southwestern region*, and thus, in the second option, the result is for Russia, not for Bulgaria. The latter passage pushes the probability mass to an option that talks about transportation (D), which is incorrect. Fortunately, the forth context has an almost full term overlap with the correct answer (B), and thus gets very high probability assigned to it: 72%.

The fourth question, from the history domain, asks to point out a missing concept, but the query is dominated by the question, and especially by underscores, leading to a single hit, counting only symbols, without any words. As expected, the model assigned uniform probability to all classes.

The last question, a history quiz, is a factoid one, and it lacks a reasoning component, unlike the previous examples. The query returned a single direct match. The retrieved passage contains the correct answer exactly: option *Bogoridi* (C). Thereby, the comprehension model assigns to it a very high probability of 68%.

#docs	Overall	biology-12th	philosophy-12th	geography-12th	history-12th	history-quiz
			Random			
0	24.89	26.09	24.44	24.18	25.87	24.03
			Window Small			
			title.bulgarian, passage.bulgarian			
1	39.95	40.27	40.63	34.97	42.99	41.99
2	40.22	40.27	40.63	35.95	42.62	42.72
5	40.22	38.90	40.63	38.07	41.51	42.48
10	38.66	40.50	39.84	35.46	39.30	38.83
20	36.84	37.53	39.05	33.82	38.75	34.71
			title.bulgarian, passage.ngram			
1	28.94	29.06	32.06	27.29	27.49	28.40
2	29.09	29.06	33.33	25.00	28.78	29.13
5	29.05	27.46	32.06	26.63	30.63	27.67
10	29.62	29.06	32.54	26.96	30.07	29.13
20	29.43	31.81	32.70	26.63	28.60	27.18
			title.bulgarian, passage.ngram, passage, passage.bulgarian			
1	38.32	38.22	40.00	34.48	39.48	40.05
2	39.08	37.07	40.32	34.48	40.59	44.17
5	39.35	40.96	39.84	34.64	41.33	41.26
10	38.63	40.50	40.63	33.50	40.41	38.83
20	36.54	38.67	37.94	31.37	37.45	38.59
			passage.ngram, passage, passage.bulgarian^2			
1	39.69	40.27	40.63	35.13	42.07	41.26
2	40.26	39.82	40.95	35.95	42.62	42.96
5	39.57	39.59	39.37	37.25	40.96	41.50
10	38.70	41.19	39.52	35.78	39.30	38.35
20	37.14	39.36	37.78	35.29	38.38	34.95
			title.bulgarian^2, passage.ngram, passage, passage.bulgarian^2			
1	39.84	40.27	40.79	35.13	42.25	41.75
2	40.30	40.27	40.63	36.11	42.80	42.72
5	40.26	39.13	40.63	38.40	41.14	42.48
10	38.74	40.50	39.68	35.62	39.48	39.08
20	37.07	37.76	39.05	34.64	38.56	34.95
			Window Big			
			title.bulgarian^2, passage.ngram, passage, passage.bulgarian^2			
1	31.22	28.38	33.97	29.41	30.81	33.25
2	33.12	31.58	37.46	31.21	33.95	29.85
5	36.04	35.70	38.10	33.82	37.82	34.22
10	36.54	37.30	36.03	33.99	39.30	36.65
20	35.62	34.55	39.68	31.05	38.38	33.74
			Paragraph			
			title.bulgarian^2, passage.ngram, passage, passage.bulgarian^2			
1	41.82	41.42	42.06	38.07	40.96	48.54
2	**42.23**	42.56	**43.17**	35.62	**42.99**	**49.27**
5	41.59	**43.25**	40.32	**38.73**	40.04	48.06
10	39.46	40.96	38.41	36.93	39.85	42.72
20	37.52	39.13	37.62	34.64	38.56	38.59
			Slavic BERT			
1	33.19	30.89	33.17	28.76	32.29	43.45
2	33.27	31.58	31.90	31.21	35.24	37.62
5	31.14	30.21	30.16	29.25	31.00	36.65
10	30.42	29.29	29.68	29.74	31.92	31.80
20	29.66	28.60	29.37	28.43	32.10	29.85

Table 5: Evaluation results for the Bulgarian multiple-choice reading comprehension task: comparison of various indexing and query strategies.

Context	Pr_A	Pr_B	Pr_C	Pr_D
✓ Q: The thick coat of mammals in winter is an example of:				
1) The hair cover is a rare and rough bristle. In winter, soft and dense hair develops between them. Color ranges from dark brown to gray, individually and geographically diverse	0.19	0.19	0.15	0.47
✗ Q: According to relativism in ethics:				
1) Moral relativism	0.45	0.24	0.10	0.21
2) In ethics, relativism is opposed to absolutism. Whilst absolutism asserts the belief that there are universal ethical standards that are inflexible and absolute, relativism claims that ethical norms vary and differ from age to age and in different cultures and situations. It can also be called epistemological relativism - a denial of absolute standards of truth evaluation.	0.28	0.41	0.09	0.22
✓ Q: Which of the assertions about the economic specialization of the Southwest region is true?				
1) Geographic and soil-climatic conditions are blessed for the development and cultivation of oil-bearing rose and other essential oil crops.	0.12	0.52	0.28	0.08
2) Kirov has an airport of regional importance. Kirov is connected with rail transport with the cities of the Transsiberian highway (Moscow and Vladivostok).	0.14	0.27	0.06	0.53
3) Dulovo has always been and remains the center of an agricultural area, famous for its grain production. The industrial sectors that still find their way into the city's economy are primarily related to the primary processing of agricultural produce. There is also the seamless production that evolved into small businesses with relatively limited economic significance.	0.25	0.05	0.67	0.03
4) In the glacial valleys and cirques and around the lakes in the highlands of Rila and Pirin, there are marshes and narrow-range glaciers (overlaps).	0.10	0.72	0.08	0.10
❓ Q: Point out the concept that is missed in the text of the Turnovo Constitution: . . .				
1) _____	0.26	0.26	0.26	0.22
✓ Q: Sofroniy Vrachanski sets up a genre that plays a big role in the history of the Bulgarian Revival. What is his name?				
1) Bogoridi is a Bulgarian Chorbadji genus from Kotel. Its founder is Bishop Sofronius Vrachanski (1739-1813). His descendants are:	0.06	0.16	0.68	0.10

Table 6: Retrieved unique top-1 contexts for the example questions in Table 2. The passages are retrieved using queries formed by concatenating a question with an answer option.