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

Momchil Hardalov[1]   Ivan Koychev[1]   Preslav Nakov[2]

[1]FMI, Sofia University „St. Kliment Ohridski", Sofia, Bulgaria,
{hardalov,koychev}@fmi.uni-sofia.bg

[2]Qatar Computing Research Institute, HBKU, Doha, Qatar
pnakov@qf.org.qa

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

SOFIA UNIVERSITY
ST. KLIMENT OHRIDSKI

Overview

Task Definition

Context: The official language of Germany is Standard German, with over 95 percent of the country speaking Standard German or German dialects as their first language. [1]

Q: *What language do people speak in Germany?*

A French

B Russian

C German

[1] https://en.wikipedia.org/wiki/Languages_of_Germany
https://quizzykid.com/quiz/general-knowledge-quiz-with-answers-multiple-choice/

Task Definition

Context: The official language of Germany is Standard German, with over 95 percent of the country speaking Standard German or German dialects as their first language. [1]

Q: *What language do people speak in Germany?*

A French

B Russian

C German

[1] `https://en.wikipedia.org/wiki/Languages_of_Germany`
https://quizzykid.com/quiz/general-knowledge-quiz-with-answers-multiple-choice/

Task Definition

Context: The tomato is the edible, often red, berry of the plant Solanum lycopersicum, commonly known as a tomato plant. [1]

Q: *What is color is the tomato?*

A Red

B Yellow

C White

[1] https://en.wikipedia.org/wiki/Tomato
https://quizzykid.com/quiz/general-knowledge-quiz-with-answers-multiple-choice/

Task Definition

Context: The tomato is the edible, often red, berry of the plant Solanum lycopersicum, commonly known as a tomato plant. [1]

Q: *What is color is the tomato?*

- A Red
- B Yellow
- C White

[1] `https://en.wikipedia.org/wiki/Tomato`
https://quizzykid.com/quiz/general-knowledge-quiz-with-answers-multiple-choice/

Task Definition

This is not that hard, right?

Task Definition

Context: Leur couleur, d'abord verdâtre, tourne généralement au rouge à maturité... [1]

Q: *De quelle couleur est une tomate?*

A Rouge

B Jaune

C Blanche

[1] https://fr.wikipedia.org/wiki/Tomate

Task Definition

Context: Leur couleur, d'abord verdâtre, tourne généralement au rouge à maturité… [1]

Q: *De quelle couleur est une tomate?*

A Rouge

B Jaune

C Blanche

[1] https://fr.wikipedia.org/wiki/Tomate

Task Definition

Context: Плодът е съестен, ярко оцветен (обикновено червен от пигмента ликопен) месест семков плод... [1]

Q: *Какъв цвят е доматът?*

A Червен

B Жълт

C Бял

[1] https://bg.wikipedia.org/wiki/%D0%94%D0%BE%D0%BC%D0%B0%D1%82

Task Definition

Context: Плодът е съестен, ярко оцветен (обикновено червен от пигмента ликопен) месест семков плод...[1]

Q: *Какъв цвят е доматът?*

- A Червен
- B Жълт
- C Бял

[1] https://bg.wikipedia.org/wiki/%D0%94%D0%BE%D0%BC%D0%B0%D1%82

Still doable?

Overview

What data is there, and how is it different?

Data & Preprocessing

Wh ent?

Data & Preprocessing

- A lot of work for English:

Data & Preprocessing

- A lot of work for English:
  - Extractive RC (MS MARCO, NewsQA, TriviaQA, SQuAD, CoQA) [Nguyen et al., 2016, Trischler et al., 2017, Joshi et al., 2017, Rajpurkar et al., 2018, Reddy et al., 2019]

Data & Preprocessing

- A lot of work for English:
  - Extractive RC (MS MARCO, NewsQA, TriviaQA, SQuAD, CoQA) [Nguyen et al., 2016, Trischler et al., 2017, Joshi et al., 2017, Rajpurkar et al., 2018, Reddy et al., 2019]
  - Non-extractive RC (MCTest, **RACE**, ARC, OpenBookQA, DREAM) [Richardson et al., 2013, Lai et al., 2017, Clark et al., 2018, Mihaylov et al., 2018, Sun et al., 2019a]

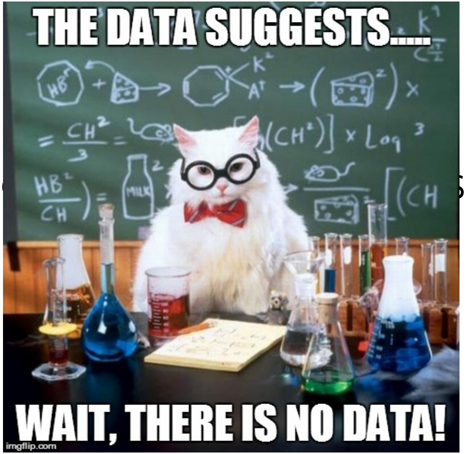
Data & Preprocessing

- A lot of work for English:
  - Extractive RC (MS MARCO, NewsQA, TriviaQA, SQuAD, CoQA) [Nguyen et al., 2016, Trischler et al., 2017, Joshi et al., 2017, Rajpurkar et al., 2018, Reddy et al., 2019]
  - Non-extractive RC (MCTest, **RACE**, ARC, OpenBookQA, DREAM) [Richardson et al., 2013, Lai et al., 2017, Clark et al., 2018, Mihaylov et al., 2018, Sun et al., 2019a]
- We chose RACE dataset for the English training [Lai et al., 2017]
  - Non-extractive multiple-choice type with context passages
  - Designed by educational experts
  - Expected to be well-structured and error-free [Sun et al., 2019a]

What about non-English datasets?

Data & Preprocessing

What ab



sets?

Data & Preprocessing

- No suitable dataset available (too few examples, different domains, not multiple-choice, etc.)

Data & Preprocessing

- No suitable dataset available (too few examples, different domains, not multiple-choice, etc.)
- We have built our own dataset for Bulgarian (2,633 questions, no contexts)

Data & Preprocessing

- No suitable dataset available (too few examples, different domains, not multiple-choice, etc.)
- We have built our own dataset for Bulgarian (2,633 questions, no contexts)
- Two question categories:
  - Online History Quizzes (Easier)
  - 12th Grade Matriculation Exam (Hard)

Data & Preprocessing

- No suitable dataset available (too few examples, different domains, not multiple-choice, etc.)
- We have built our own dataset for Bulgarian (2,633 questions, no contexts)
- Two question categories:
  - Online History Quizzes (Easier)
  - 12th Grade Matriculation Exam (Hard)
- Manually filtered out questions:
  - with non-textual content (i.e., pictures, paintings, drawings, etc.)
  - ordering questions (i.e., order the historical events)
  - questions involving calculations (i.e., how much X we need to add to Y to arrive at Z)

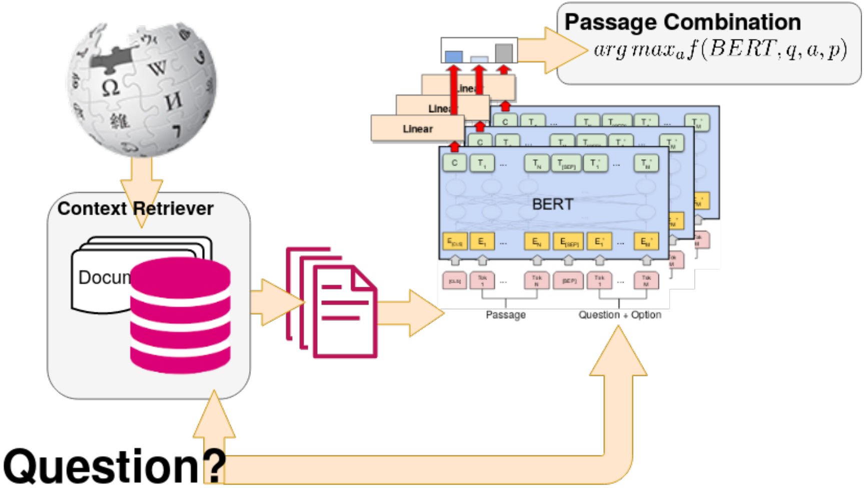
Data statistics

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

Overview

Model Overview



- Answer 1
- Answer 2
- ...
- Answer N

Context Retriever

- Dumps for the entire Wikipage[1]

[1] http://dumps.wikimedia.org/

Context Retriever

- Dumps for the entire Wikipage[1]
- Removing links, HTML tags, tables, etc.

[1] http://dumps.wikimedia.org/

Context Retriever

- Dumps for the entire Wikipage[1]
- Removing links, HTML tags, tables, etc.
- Two document splitting strategies:
  - paragraph
  - sliding window

[1] http://dumps.wikimedia.org/

Context Retriever

- Dumps for the entire Wikipage[1]
- Removing links, HTML tags, tables, etc.
- Two document splitting strategies:
  - paragraph
  - sliding window
- Query is formed from a question and possible answers

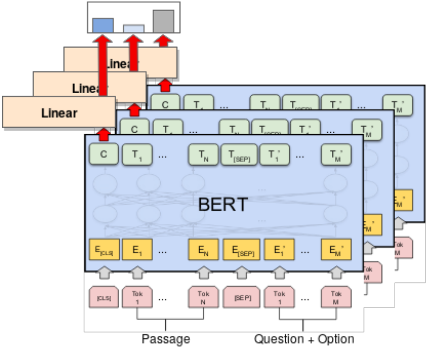
[1] http://dumps.wikimedia.org/

Context Retriever

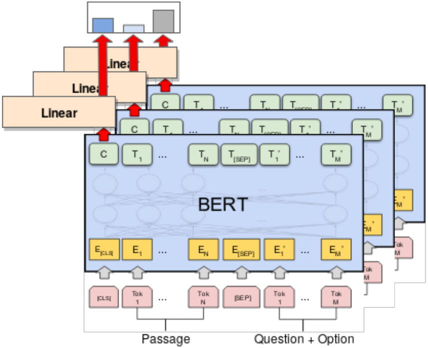
- Dumps for the entire Wikipage[1]
- Removing links, HTML tags, tables, etc.
- Two document splitting strategies:
  - paragraph
  - sliding window
- Query is formed from a question and possible answers
- Matching with cosine similarity and BM25 (Improved TF.IDF) [Robertson and Zaragoza, 2009]

[1] http://dumps.wikimedia.org/

BERT for Reading Comprehension

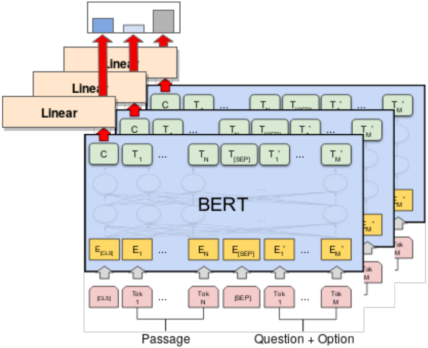


BERT for Reading Comprehension



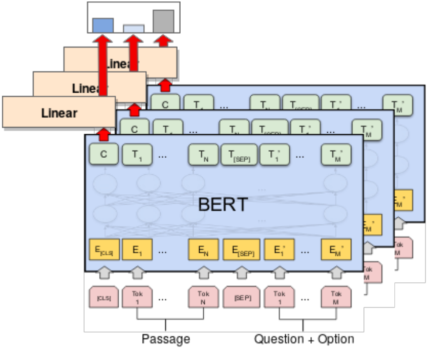
- Model input: *[CLS] Passage [SEP] Question + Option [SEP]*

BERT for Reading Comprehension



- Model input: *[CLS] Passage [SEP] Question + Option [SEP]*
- Task-specific parameter vector L, $L \in \mathbb{R}^H$, where H is the hidden size of the model

BERT for Reading Comprehension



- Model input: *[CLS] Passage [SEP] Question + Option [SEP]*
- Task-specific parameter vector L, $L \in \mathbb{R}^H$, where H is the hidden size of the model
- Maximizing the log-probability of the correct answer

Answer Selection Strategies

Why do we need something special? Why not just the $\arg\max$?

Answer Selection Strategies

Why do we need something special? Why not just the $\arg\max$?

- The first hit is not always the best

Answer Selection Strategies

Why do we need something special? Why not just the $\arg\max$?

- The first hit is not always the best
- The retriever can be extremely sensitive to the question formulation

Answer Selection Strategies

Why do we need something special? Why not just the $\arg\max$?

- The first hit is not always the best
- The retriever can be extremely sensitive to the question formulation

How do we fix it?

Answer Selection Strategies

Why do we need something special? Why not just the $\arg\max$?

- The first hit is not always the best
- The retriever can be extremely sensitive to the question formulation

How do we fix it? – *Use multiple documents to obtain better "correctness distribution"*

Answer Selection Strategies

Why do we need something special? Why not just the arg max?

- The first hit is not always the best
- The retriever can be extremely sensitive to the question formulation

How do we fix it? – *Use multiple documents to obtain better "correctness distribution"*

Strategy Formalization

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

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

Answer Selection Strategies

Why do we need something special? Why not just the arg max?

- The first hit is not always the best
- The retriever can be extremely sensitive to the question formulation

How do we fix it? – *Use multiple documents to obtain better "correctness distribution"*

Strategy Formalization

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

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

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

Overview

BERT Fine-Tuning

Fine-tuning on English multiple-choice questions from the RACE dataset.

On top of two flavours of BERT:

BERT Fine-Tuning

Fine-tuning on English multiple-choice questions from the RACE dataset.

On top of two flavours of BERT:

Multilingual

- $BERT_{base}$ Cased (12-layers, 768-hidden, 12-heads)
- Pre-trained on 104 languages

BERT Fine-Tuning

Fine-tuning on English multiple-choice questions from the RACE dataset.

On top of two flavours of BERT:

Multilingual vs. *Slavic*

Multilingual

- *BERT$_{base}$* Cased (12-layers, 768-hidden, 12-heads)
- Pre-trained on 104 languages

Slavic

- Additional training on Slavic languages (BG, CZ, PL, RU)
- News + Wikipedia articles

Results on the English Task

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

Zero-Shot Transfer to Bulgarian

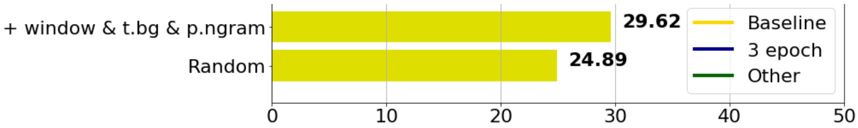


Figure: Accuracy on the Bulgarian testset

Zero-Shot Transfer to Bulgarian

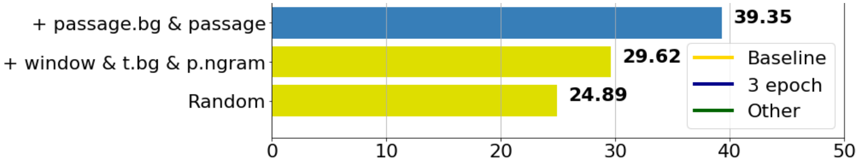


Figure: Accuracy on the Bulgarian testset

Zero-Shot Transfer to Bulgarian

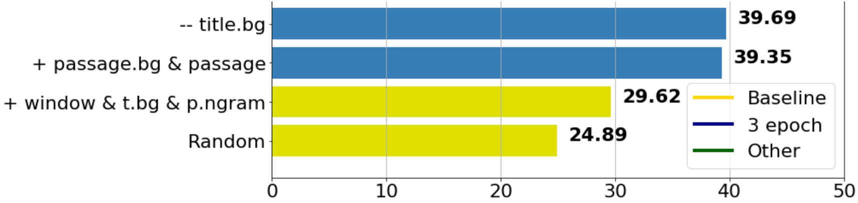


Figure: Accuracy on the Bulgarian testset

Zero-Shot Transfer to Bulgarian

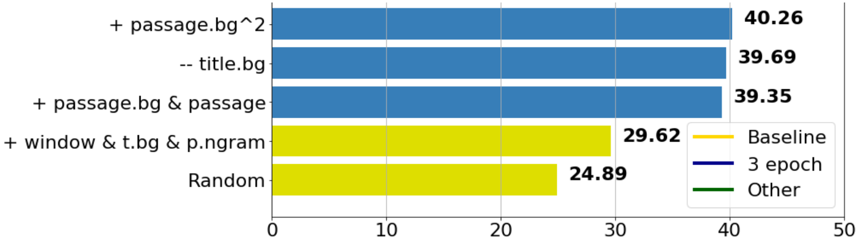


Figure: Accuracy on the Bulgarian testset

Zero-Shot Transfer to Bulgarian

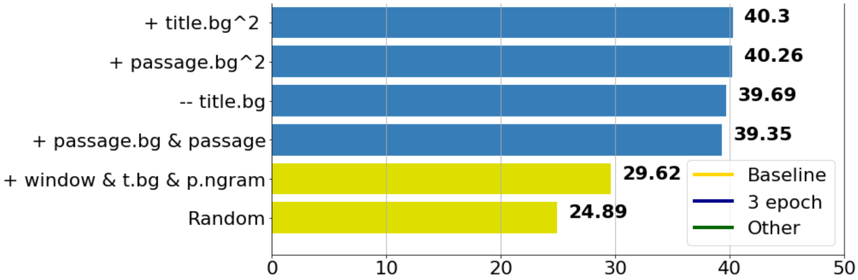


Figure: Accuracy on the Bulgarian testset

Zero-Shot Transfer to Bulgarian

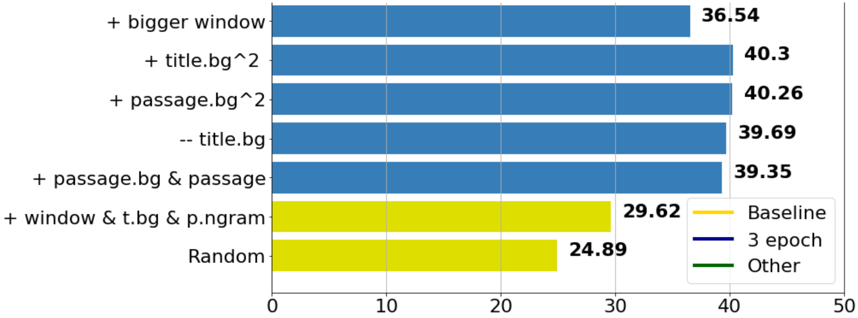


Figure: Accuracy on the Bulgarian testset

Zero-Shot Transfer to Bulgarian

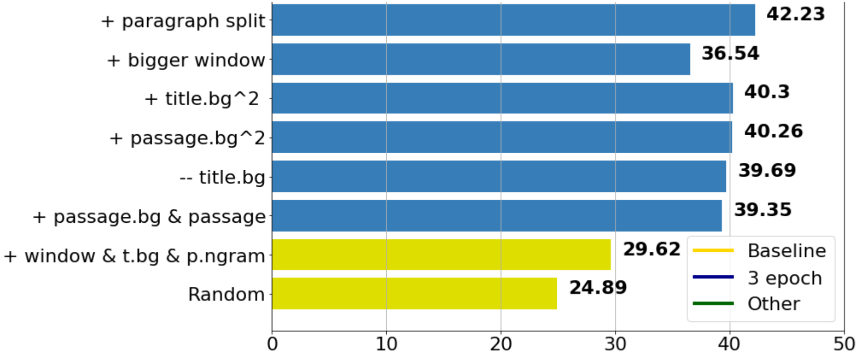


Figure: Accuracy on the Bulgarian testset

Zero-Shot Transfer to Bulgarian

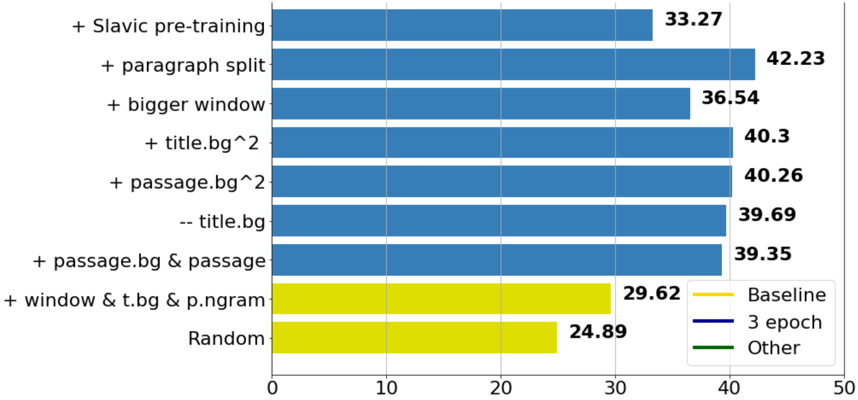


Figure: Accuracy on the Bulgarian testset

Zero-Shot Transfer to Bulgarian

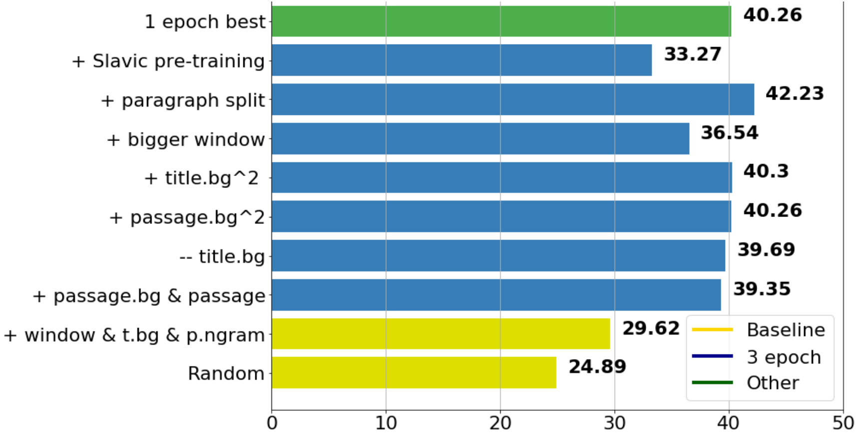


Figure: Accuracy on the Bulgarian testset

Zero-Shot Transfer to Bulgarian

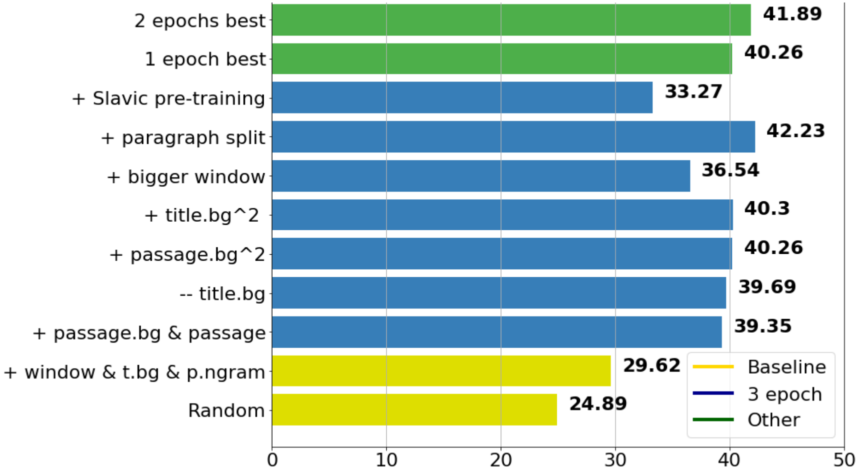
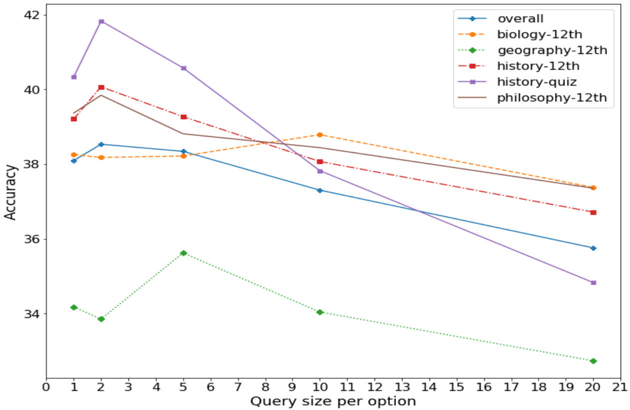


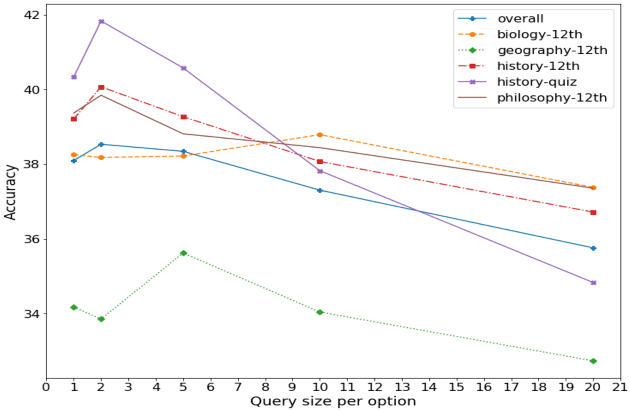
Figure: Accuracy on the Bulgarian testset

Impact of the Query Result Size



We experiment with query sizes $S_q \in \{1, 2, 5, 10, 20\}$

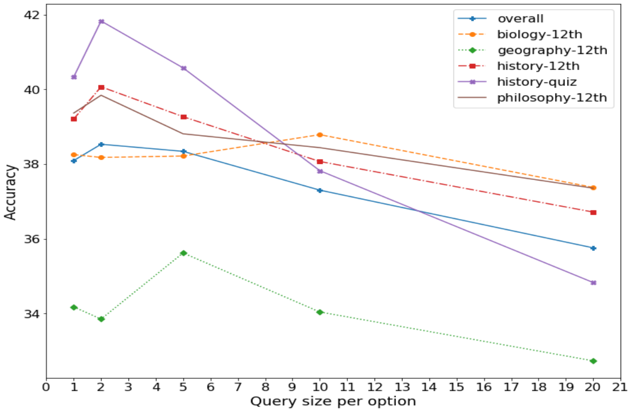
Impact of the Query Result Size



We experiment with query sizes $S_q \in \{1, 2, 5, 10, 20\}$

- Results are averaged over all experiments

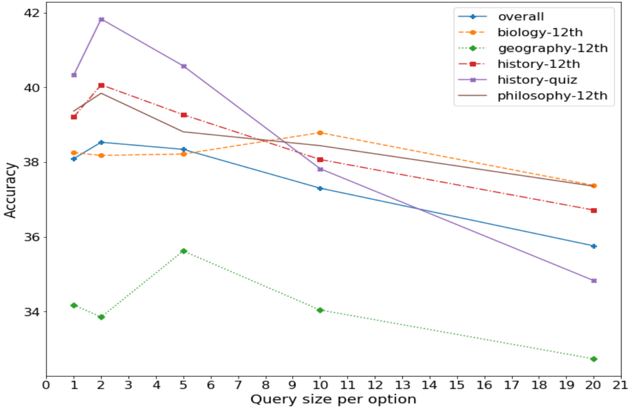
Impact of the Query Result Size



We experiment with query sizes $S_q \in \{1, 2, 5, 10, 20\}$

- Results are averaged over all experiments
- Up to $S_q \times \#Options$ hits

Impact of the Query Result Size



We experiment with query sizes $S_q \in \{1, 2, 5, 10, 20\}$

- Results are averaged over all experiments
- Up to $S_q \times \#Options$ hits
- Duplicates are merged

Examples I

Retrieved Contexts:

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

Question	$Pr_{ctx(1)}$
✓ Q: The thick coat of mammals in winter is an example of:	
A. physiological adaptation	0.19
B. behavioral adaptation	0.19
C. genetic adaptation	0.15
D. morphological adaptation	0.47

Examples II

Retrieved Contexts:

1. Moral relativism
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.

Question	$Pr_{ctx(1)}$	$Pr_{ctx(2)}$
✗ Q: According to relativism in ethics:		
A. there is only one moral law that is valid for all	0.45	0.28
B. there is no absolute good and evil	0.24	0.41
C. people are evil by nature	0.09	0.10
D. there is only good, and the evil is seeming	0.21	0.22

Overview

Literature Review I

Machine Reading Comprehension

- Usage of external knowledge:
  - Wikipedia for answering open-domain questions [Chen et al., 2017a]
  - Entity discovery and linking [Pan et al., 2018]
  - Semi-automatically constructed knowledge base [Clark et al., 2016]

Literature Review I

Machine Reading Comprehension

- Usage of external knowledge:
  - Wikipedia for answering open-domain questions [Chen et al., 2017a]
  - Entity discovery and linking [Pan et al., 2018]
  - Semi-automatically constructed knowledge base [Clark et al., 2016]
- Question/Answer reformulation:
  - Finding essential terms, and query reformulation [Ni et al., 2019]
  - Conversion to declarative sentences and linguistic units [Simov et al., 2012]

Literature Review I

Machine Reading Comprehension

- Usage of external knowledge:
  - Wikipedia for answering open-domain questions [Chen et al., 2017a]
  - Entity discovery and linking [Pan et al., 2018]
  - Semi-automatically constructed knowledge base [Clark et al., 2016]
- Question/Answer reformulation:
  - Finding essential terms, and query reformulation [Ni et al., 2019]
  - Conversion to declarative sentences and linguistic units [Simov et al., 2012]
- Application of reading strategies [Sun et al., 2019b]

Literature Review II

(Zero-Shot) Multilingual Models

- Fine-tuned multilingual language models BERT [Devlin et al., 2019], and XLM [Lample and Conneau, 2019]

Literature Review II

(Zero-Shot) Multilingual Models

- Fine-tuned multilingual language models BERT [Devlin et al., 2019], and XLM [Lample and Conneau, 2019]
- Shared model:
  - Seq2Seq with a special token for each language [Johnson et al., 2017]
  - Many-to-one language training with a shared attention layer [Firat et al., 2016]
  - Many-to-many languages with a single Transformer model [Aharoni et al., 2019]

Literature Review II

(Zero-Shot) Multilingual Models

- Fine-tuned multilingual language models BERT [Devlin et al., 2019], and XLM [Lample and Conneau, 2019]
- Shared model:
  - Seq2Seq with a special token for each language [Johnson et al., 2017]
  - Many-to-one language training with a shared attention layer [Firat et al., 2016]
  - Many-to-many languages with a single Transformer model [Aharoni et al., 2019]
- Pivot-language approaches:
  - Student-teacher framework for NMT [Chen et al., 2017b]
  - Translation and soft-alignment for MRC [Asai et al., 2018]

Overview

Conclusions & Future Work

Conclusions

- Collected a corpus in Bulgarian with 2,633 questions [2]

[2] Dataset and source code: `http://github.com/mhardalov/bg-reason-BERT`

Conclusions & Future Work

Conclusions

- Collected a corpus in Bulgarian with 2,633 questions [2]
- Designed a general-purpose pipeline for multiple-choice MRC

[2] Dataset and source code: `http://github.com/mhardalov/bg-reason-BERT`

Conclusions & Future Work

Conclusions

- Collected a corpus in Bulgarian with 2,633 questions [2]
- Designed a general-purpose pipeline for multiple-choice MRC
- Studied the effectiveness of zero-shot transferred model from English to Bulgarian

[2] Dataset and source code: `http://github.com/mhardalov/bg-reason-BERT`

Conclusions & Future Work

Conclusions

- Collected a corpus in Bulgarian with 2,633 questions [2]
- Designed a general-purpose pipeline for multiple-choice MRC
- Studied the effectiveness of zero-shot transferred model from English to Bulgarian
- Achieved 42.24% accuracy (well above the baseline of 24.89%)

[2] Dataset and source code: `http://github.com/mhardalov/bg-reason-BERT`

Conclusions & Future Work

Conclusions

- Collected a corpus in Bulgarian with 2,633 questions [2]
- Designed a general-purpose pipeline for multiple-choice MRC
- Studied the effectiveness of zero-shot transferred model from English to Bulgarian
- Achieved 42.24% accuracy (well above the baseline of 24.89%)

[2] Dataset and source code: `http://github.com/mhardalov/bg-reason-BERT`

Conclusions & Future Work

Conclusions

- Collected a corpus in Bulgarian with 2,633 questions [2]
- Designed a general-purpose pipeline for multiple-choice MRC
- Studied the effectiveness of zero-shot transferred model from English to Bulgarian
- Achieved 42.24% accuracy (well above the baseline of 24.89%)

Future Work

- Reading strategies [Sun et al., 2019b]
- Linked entities [Pan et al., 2018]
- Reformulation of questions and passages [Simov et al., 2012, Clark et al., 2016, Ni et al., 2019]
- Re-ranking of documents [Nogueira and Cho, 2019]

[2] Dataset and source code: `http://github.com/mhardalov/bg-reason-BERT`

References I

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

Asai, A., Eriguchi, A., Hashimoto, K., and Tsuruoka, Y. (2018).
Multilingual extractive reading comprehension by runtime machine translation.
arXiv preprint arXiv:1809.03275.

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

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

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. (2018).
Think you have solved question answering? Try ARC, the AI2 reasoning challenge.
arXiv preprint arXiv:1803.05457.

Clark, P., Etzioni, O., Khot, T., Sabharwal, A., Tafjord, O., Turney, P., and Khashabi, D. (2016).
Combining retrieval, statistics, and inference to answer elementary science questions.
In *Proceedings of the 13th AAAI Conference on Artificial Intelligence*, AAAI '16, pages 2580–2586, Phoenix, AZ, USA.

References II

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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*, NAACL-HLT '19, pages 4171–4186, Minneapolis, MN, USA.

Firat, O., Sankaran, B., Al-Onaizan, Y., Yarman Vural, F. T., and Cho, K. (2016).
Zero-resource translation with multi-lingual neural machine translation.
In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '16, pages 268–277, Austin, TX, USA.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017).
Google's multilingual neural machine translation system: Enabling zero-shot translation.
Transactions of the Association for Computational Linguistics, 5:339–351.

Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. (2017).
TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension.
In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, ACL '17, pages 1601–1611, Vancouver, Canada.

Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. (2017).
RACE: Large-scale ReAding comprehension dataset from examinations.
In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 785–794, Copenhagen, Denmark.

Lample, G. and Conneau, A. (2019).
Cross-lingual language model pretraining.
arXiv preprint arXiv:1901.07291.

References III

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. (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*, EMNLP '18, pages 2381–2391, Brussels, Belgium.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016).
MS MARCO: A human generated machine reading comprehension dataset.
In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*, CoCo@NIPS '16, Barcelona, Spain.

Ni, J., Zhu, C., Chen, W., and McAuley, J. (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*, NAACL-HLT '19, pages 335–344, Minneapolis, MN, USA.

Nogueira, R. and Cho, K. (2019).
Passage re-ranking with BERT.
arXiv preprint arXiv:1901.04085.

Pan, X., Sun, K., Yu, D., Ji, H., and Yu, D. (2018).
Improving question answering with external knowledge.
arXiv preprint:1902.00993.

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

References IV

Reddy, S., Chen, D., and Manning, C. D. (2019).
CoQA: A conversational question answering challenge.
Transactions of the Association for Computational Linguistics, 7:249–266.

Richardson, M., Burges, C. J., and Renshaw, E. (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*, EMNLP '13, pages 193–203, Seattle, WA, USA.

Robertson, S. and Zaragoza, H. (2009).
The probabilistic relevance framework: BM25 and beyond.
Found. Trends Inf. Retr., 3(4):333–389.

Simov, K. I., Osenova, P., Georgiev, G., Zhikov, V., and Tolosi, L. (2012).
Bulgarian question answering for machine reading.
In *CLEF Working Note Papers*, Rome, Italy.

Sun, K., Yu, D., Chen, J., Yu, D., Choi, Y., and Cardie, C. (2019a).
DREAM: A challenge data set and models for dialogue-based reading comprehension.
Transactions of the Association for Computational Linguistics, 7:217–231.

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

Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K. (2017).
NewsQA: A machine comprehension dataset.
In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, RepL4NLP '19, pages 191–200, Vancouver, Canada.

Results

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

Results per category

#docs	Overall	biology-12th	philosophy-12th	geography-12th	history-12th	history-quiz
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: Evaluation results for the Bulgarian multiple-choice reading comprehension task: comparison of various indexing and query strategies.