

# Cross-lingual Training for Multiple-Choice Question Answering

---

**Guillermo Echegoyen Blanco**

Álvaro Rodrigo

Anselmo Peñas

{gblanco, alvarory, anselmo} **at** lsi.uned.es

NLP & IR Group

National Distance Education University (UNED)

UNED

2020-09-17

Cross-lingual Training for Multiple-Choice Question  
Answering

Cross-lingual Training for  
Multiple-Choice Question Answering

Guillermo Echegoyen Blanco  
Álvaro Rodrigo  
Anselmo Peñas  
{gblanco, alvarory, anselmo} at lsi.uned.es

NLP & IR Group  
National Distance Education University (UNED)

1. Introduction

2. Problem Statement

3. Experiments

4. Results

5. Conclusions & Future Work

6. References

2020-09-17

## Cross-lingual Training for Multiple-Choice Question Answering

└ Overview

1. Introduction

2. Problem Statement

3. Experiments

4. Results

5. Conclusions & Future Work

6. References

2020-09-17

Cross-lingual Training for Multiple-Choice Question  
Answering  
└ Introduction

Introduction

# Introduction

---

## Multiple-Choice Question Answering

**Def:** Given a supporting text, a question and a set of possible answers, choose the correct one.

**Example** (taken from RACE (Lai et al. 2017))

**Evidence:** . . . Many people optimistically thought industry awards for better equipment would stimulate the production of quieter appliances. It was even suggested that noise from building sites could be alleviated . . .

**Question:** *What was the author's attitude towards the industry awards for quieter?*

**Options:** A. **suspicious** C. enthusiastic D. indifferent

2020-09-17

## Cross-lingual Training for Multiple-Choice Question Answering

- Introduction

- Introduction

### Multiple-Choice Question Answering

**Def:** Given a supporting text, a question and a set of possible answers, choose the correct one.

**Example** (taken from RACE (Lai et al. 2017))

**Evidence:** . . . Many people optimistically thought industry awards for better equipment would stimulate the production of quieter appliances. It was even suggested that noise from building sites could be alleviated . . .

**Question:** *What was the author's attitude towards the industry awards for quieter?*

**Options:** A. **suspicious** C. enthusiastic D. indifferent

## Multiple-Choice Question Answering

- Measure reading comprehension in humans.
- Collections are usually extracted from exams for humans.
- Many real world exams are private.
- The majority of dataset are in English.

2020-09-17

## Cross-lingual Training for Multiple-Choice Question Answering

### └ Introduction

### └ Introduction

- Private exams are not suitable for research

#### Multiple-Choice Question Answering

- Measure reading comprehension in humans.
- Collections are usually extracted from exams for humans.
- Many real world exams are private.
- The majority of dataset are in English.

## Motivation

- Scarce non-English datasets.
- Non-English datasets are usually small.

2020-09-17

## Cross-lingual Training for Multiple-Choice Question Answering

### └ Introduction

### └ Introduction

- Not enough data to train in non-English collections

Introduction

#### Motivation

- Scarce non-English datasets.
- Non-English datasets are usually small.

## Research Questions

- How to zero-shot transfer from a big MC-QA collection to another one?
- Can we zero-shot transfer to another collection in a different language?
- Harder exams for humans are so for machines too?

#### Research Questions

- How to zero-shot transfer from a big MC-QA collection to another one?
- Can we zero-shot transfer to another collection in a different language?
- Harder exams for humans are so for machines too?

2020-09-17

Cross-lingual Training for Multiple-Choice Question  
Answering  
└ Problem Statement

Problem Statement

## Problem Statement

---



## Datasets

### RACE (Lai et al. 2017)

- Chinese schools exams
- > 97K Questions
- English (monolingual)

### Entrance Exams (Rodrigo et al. 2018)

- University access in Japan
- $\approx$  200 Questions
- 6 languages (multilingual)

2020-09-17

## Cross-lingual Training for Multiple-Choice Question Answering

### └ Problem Statement

### └ Problem Statement

- RACE:
  1. Divided into two collections: middle and high school.
- EE:
  1.  $\approx$  500 times smaller than RACE.
  2. Not suitable for fine tuning.
  3. Translated to Spanish, Italian, French, Russian, and (just one edition) German.
  4. Exams from three different years
- RACE are pre University exams, EE are exams at University level.

Problem Statement

#### Datasets

##### RACE (Lai et al. 2017)

- Chinese schools exams
- > 97K Questions
- English (monolingual)

##### Entrance Exams (Rodrigo et al. 2018)

- University access in Japan
- $\approx$  200 Questions
- 6 languages (multilingual)

# Problem Statement

## Approach

Not enough data on Entrance Exams for training:

- Train over RACE
- Evaluate over Entrance Exams

2020-09-17

Cross-lingual Training for Multiple-Choice Question Answering

└ Problem Statement

└ Problem Statement

Problem Statement

### Approach

Not enough data on Entrance Exams for training:

- Train over RACE
- Evaluate over Entrance Exams

2020-09-17

Cross-lingual Training for Multiple-Choice Question  
Answering  
└ Experiments

Experiments

## Experiments

---

## Method

- No hyper-parameters search.
- Fine-tune each model over RACE.
- Test each model over RACE.
- Test each model over Entrance Exams in all languages and all years

2020-09-17

## Cross-lingual Training for Multiple-Choice Question Answering

- └ Experiments

- └ Experiments

### Method

- No hyper-parameters search.
- Fine-tune each model over RACE.
- Test each model over RACE.
- Test each model over Entrance Exams in all languages and all years

## Models

- BERT-base
- Multi BERT-base

## Baselines

- Random
- Longest answer (Rogers et al. 2020)

2020-09-17

## Cross-lingual Training for Multiple-Choice Question Answering

- └ Experiments

- └ Experiments

Experiments

| UNED

### Models

- BERT-base
- Multi BERT-base

### Baselines

- Random
- Longest answer (Rogers et al. 2020)

2020-09-17

Cross-lingual Training for Multiple-Choice Question  
Answering  
└ Results

Results

## Results

# Results

Dataset	BERT	MultiBERT	Random	Longest
RACE Mid	0.5265	<b>0.6114</b>	0.2500	0.3078
RACE High	0.4774	<b>0.5031</b>	0.2500	0.3059
RACE All	0.4917	<b>0.5347</b>	0.2500	0.3059
EE English	0.4921	<b>0.4974</b>	0.2500	0.2304
EE Spanish	0.3665	<b>0.4503</b>	0.2500	0.2932
EE Italian	0.2880	<b>0.4293</b>	0.2500	0.2775
EE French	0.3037	<b>0.4346</b>	0.2500	0.2565
EE Russian	0.2618	<b>0.3403</b>	0.2500	0.2723
EE German**	0.3708	<b>0.4494</b>	0.2500	0.2584

2020-09-17

## Cross-lingual Training for Multiple-Choice Question Answering

└ Results

└ Results

Dataset	BERT	MultiBERT	Random	Longest
RACE Mid	0.5265	<b>0.6114</b>	0.2500	0.3078
RACE High	0.4774	<b>0.5031</b>	0.2500	0.3059
RACE All	0.4917	<b>0.5347</b>	0.2500	0.3059
EE English	0.4921	<b>0.4974</b>	0.2500	0.2304
EE Spanish	0.3665	<b>0.4503</b>	0.2500	0.2932
EE Italian	0.2880	<b>0.4293</b>	0.2500	0.2775
EE French	0.3037	<b>0.4346</b>	0.2500	0.2565
EE Russian	0.2618	<b>0.3403</b>	0.2500	0.2723
EE German**	0.3708	<b>0.4494</b>	0.2500	0.2584

# Results

Dataset	BERT	MultiBERT	Random	Longest
RACE Mid	0.5265	<b>0.6114</b>	0.2500	0.3078
RACE High	0.4774	<b>0.5031</b>	0.2500	0.3059
RACE All	0.4917	<b>0.5347</b>	0.2500	0.3059
EE English	0.4921	<b>0.4974</b>	0.2500	0.2304
EE Spanish	0.3665	<b>0.4503</b>	0.2500	0.2932
EE Italian	0.2880	<b>0.4293</b>	0.2500	0.2775
EE French	0.3037	<b>0.4346</b>	0.2500	0.2565
EE Russian	0.2618	<b>0.3403</b>	0.2500	0.2723
EE German**	0.3708	<b>0.4494</b>	0.2500	0.2584

Difficulty affects machines too

2020-09-17

## Cross-lingual Training for Multiple-Choice Question Answering

Results

Results

Dataset	BERT	MultiBERT	Random	Longest
RACE Mid	0.5265	<b>0.6114</b>	0.2500	0.3078
RACE High	0.4774	<b>0.5031</b>	0.2500	0.3059
RACE All	0.4917	<b>0.5347</b>	0.2500	0.3059
EE English	0.4921	<b>0.4974</b>	0.2500	0.2304
EE Spanish	0.3665	<b>0.4503</b>	0.2500	0.2932
EE Italian	0.2880	<b>0.4293</b>	0.2500	0.2775
EE French	0.3037	<b>0.4346</b>	0.2500	0.2565
EE Russian	0.2618	<b>0.3403</b>	0.2500	0.2723
EE German**	0.3708	<b>0.4494</b>	0.2500	0.2584

Difficulty affects machines too



# Results

Dataset	BERT	MultiBERT	Random	Longest
RACE Mid	0.5265	<b>0.6114</b>	0.2500	0.3078
RACE High	0.4774	<b>0.5031</b>	0.2500	0.3059
RACE All	0.4917	<b>0.5347</b>	0.2500	0.3059
EE English	0.4921	<b>0.4974</b>	0.2500	0.2304
EE Spanish	0.3665	<b>0.4503</b>	0.2500	0.2932
EE Italian	0.2880	<b>0.4293</b>	0.2500	0.2775
EE French	0.3037	<b>0.4346</b>	0.2500	0.2565
EE Russian	0.2618	<b>0.3403</b>	0.2500	0.2723
EE German**	0.3708	<b>0.4494</b>	0.2500	0.2584

Pre-university graded exams results are comparable

2020-09-17

## Cross-lingual Training for Multiple-Choice Question Answering

Results

Results

Dataset	BERT	MultiBERT	Random	Longest
RACE Mid	0.5265	<b>0.6114</b>	0.2500	0.3078
RACE High	0.4774	<b>0.5031</b>	0.2500	0.3059
RACE All	0.4917	<b>0.5347</b>	0.2500	0.3059
EE English	0.4921	<b>0.4974</b>	0.2500	0.2304
EE Spanish	0.3665	<b>0.4503</b>	0.2500	0.2932
EE Italian	0.2880	<b>0.4293</b>	0.2500	0.2775
EE French	0.3037	<b>0.4346</b>	0.2500	0.2565
EE Russian	0.2618	<b>0.3403</b>	0.2500	0.2723
EE German**	0.3708	<b>0.4494</b>	0.2500	0.2584

Pre-university graded exams results are comparable

# Results

Dataset	BERT	MultiBERT	Random	Longest
RACE Mid	0.5265	<b>0.6114</b>	0.2500	0.3078
RACE High	0.4774	<b>0.5031</b>	0.2500	0.3059
RACE All	0.4917	<b>0.5347</b>	0.2500	0.3059
EE English	0.4921	<b>0.4974</b>	0.2500	0.2304
EE Spanish	0.3665	<b>0.4503</b>	0.2500	0.2932
EE Italian	0.2880	<b>0.4293</b>	0.2500	0.2775
EE French	0.3037	<b>0.4346</b>	0.2500	0.2565
EE Russian	0.2618	<b>0.3403</b>	0.2500	0.2723
EE German**	0.3708	<b>0.4494</b>	0.2500	0.2584

Best results are always in english

2020-09-17

## Cross-lingual Training for Multiple-Choice Question Answering

Results

Results

Dataset	BERT	MultiBERT	Random	Longest
RACE Mid	0.5265	<b>0.6114</b>	0.2500	0.3078
RACE High	0.4774	<b>0.5031</b>	0.2500	0.3059
RACE All	0.4917	<b>0.5347</b>	0.2500	0.3059
EE English	0.4921	<b>0.4974</b>	0.2500	0.2304
EE Spanish	0.3665	<b>0.4503</b>	0.2500	0.2932
EE Italian	0.2880	<b>0.4293</b>	0.2500	0.2775
EE French	0.3037	<b>0.4346</b>	0.2500	0.2565
EE Russian	0.2618	<b>0.3403</b>	0.2500	0.2723
EE German**	0.3708	<b>0.4494</b>	0.2500	0.2584

Best results are always in english

- Multi-BERT performs better in all scenarios
- Russian is specially difficult as language with very different semantics are not well understood nor tokenized by the model.
- \*\* German only available for one year.
- ToDo := Information about previous results
- SOTA on Entrance Exams in several languages: Spanish, Italian and German.

2020-09-17

Cross-lingual Training for Multiple-Choice Question  
Answering  
└─ Conclusions & Future Work

Conclusions & Future Work

## Conclusions & Future Work

---

## Conclusions

- Performance holds across different tasks.
- Performance holds across languages in multilingual models.
- Performance drops with difficulty for humans.

## Future Work

- Transfer knowledge learnt in one language to another one.

2020-09-17

## Cross-lingual Training for Multiple-Choice Question Answering

### └ Conclusions & Future Work

### └ Conclusions & Future Work

- When exams are more difficult for humans, they are so for machines ( $\text{Mid} < \text{High} < \text{EE}$ )
- FW: specially when languages are low-resourced.

#### Conclusions

- Performance holds across different tasks.
- Performance holds across languages in multilingual models.
- Performance drops with difficulty for humans.

#### Future Work

- Transfer knowledge learnt in one language to another one.

**Thank you!**  
**Questions?**

2020-09-17

**Cross-lingual Training for Multiple-Choice  
Question Answering**

└ **Conclusions & Future Work**

Thank you!  
Questions?

2020-09-17

Cross-lingual Training for Multiple-Choice Question  
Answering  
└─ References

References

## References

---

## References

---



Guokun Lai et al. “RACE: Large-scale ReAding Comprehension Dataset From Examinations”. In: *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings* (Apr. 2017), pp. 785–794. arXiv: 1704.04683. URL: <http://arxiv.org/abs/1704.04683>.



Guokun Lai et al. “RACE: Large-scale ReAding Comprehension Dataset From Examinations”. In: *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings* (Apr. 2017), pp. 785–794. arXiv: 1704.04683. URL: <http://arxiv.org/abs/1704.04683>.

# References ii



Alvaro Rodrigo et al. “Do systems pass university entrance exams?” In: *Information Processing & Management* 54.4 (July 2018), pp. 564–575. ISSN: 0306-4573. DOI: 10.1016/J.IPM.2018.03.002. URL: <https://www.sciencedirect.com/science/article/abs/pii/S0306457317305344>.

2020-09-17

## Cross-lingual Training for Multiple-Choice Question Answering

└ References

└ References



Alvaro Rodrigo et al. “Do systems pass university entrance exams?” In: *Information Processing & Management* 54.4 (July 2018), pp. 564–575. ISSN: 0306-4573. DOI: 10.1016/J.IPM.2018.03.002. URL: <https://www.sciencedirect.com/science/article/abs/pii/S0306457317305344>.





Anna Rogers et al. “Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05 (Apr. 2020), pp. 8722–8731. ISSN: 2374-3468. DOI: 10.1609/aaai.v34i05.6398. URL: <https://aaai.org/ojs/index.php/AAAI/article/view/6398>.

