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)

Overview

1. Introduction
2. Problem Statement
3. Results
4. Conclusions & Future Work
5. Outcomes
6. References

Introduction

Multiple-Choice Question Answering

Def: Given a supporting text, a question and a set of possible answers, choose the correct one. Commonly used to measure reading comprehension in humans.

- The majority of datasets are in English.
- Non-English datasets are usually small.
- Usually extracted from exams for humans.

ToDo := Add example?, name collections?

Introduction

Motivation

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

Problem Statement

Problem Statement

Datasets

- **RACE** (Lai et al. 2017): Collected from Chinese schools English exams. $>$ **97K questions** with, 4 possible answers each, English monolingual.
- **Entrance Exams** (Rodrigo et al. 2018): University access exams in Japan. $\approx$ **200 questions**, 4 possible answers each. Crowd-translated to 4 different languages.

Example (taken from RACE)

Evidence: ”The park is open from 8 am to 5 pm.”

Question: The park is open for __ hours a day.

Options: A. eight B. nine C. ten D. eleven

Problem Statement

Models & Baselines

- BERT-base
- Multi BERT-base
- Random
- Longest answer (Rogers et al. 2020)

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

Results

Results

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

** German only available for one year.

Conclusions & Future Work

Conclusions & Future Work

Conclusions

- Zero-shot transfer to a smaller task still holds performance in the same language.
- Can be done to a different task and language with a multilingual model.
- Performance is hampered by exams difficulty in the same way human grades do.

Future Work

- Continue exploring low-resource languages.

ToDo:= Remove Future work? Add something else?

Outcomes

Outcomes

Out main contributions are:

- SOTA on Entrance Exams in several languages.
- RACE trained BERT and Multi BERT models.

Thank you!

Questions?

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`.

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`.