Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 37-44 recibido 30-04-2020 revisado 04-06-2020 aceptado 04-06-2020

Cross-lingual Training for Multiple-Choice Question Answering

Entrenamiento Croslingüe para Búsqueda de Respuestas de Opción Múltiple

Guillermo Echegoyen, Álvaro Rodrigo, Anselmo Peñas
Universidad Nacional de Educación a Distancia (UNED)
{gblanco, alvarory, anselmo}@lsi.uned.es

Abstract: In this work we explore to what extent multilingual models can be trained for one language and applied to a different one for the task of Multiple Choice Question Answering. We employ the RACE dataset to fine-tune both a monolingual and a multilingual models and apply these models to another different collections in different languages. The results show that both monolingual and multilingual models can be zero-shot transferred to a different dataset in the same language maintaining its performance. Besides, the multilingual model still performs good when it is applied to a different target language. Additionally, we find that exams that are more difficult to humans are harder for machines too. Finally, we advance the state-of-the-art for the QA4MRE Entrance Exams dataset in several languages.
Keywords: Question Answering; Multiple-Choice Reading Comprehension; Multilinguality

Resumen: En este trabajo exploramos en qué medida los modelos multilingües pueden ser entrenados para un solo idioma y aplicados a otro diferente para la tarea de respuesta a preguntas de opción múltiple. Empleamos el conjunto de datos RACE para ajustar tanto un modelo monolingüe como multilingüe y aplicamos estos modelos a otras colecciones en idiomas diferentes. Los resultados muestran que tanto los modelos monolingües como los multilingües pueden transferirse a un conjunto de datos diferente en el mismo idioma manteniendo su rendimiento. Además, el modelo multilingüe todavía funciona bien cuando se aplica a un idioma de destino diferente. Asimismo, hemos comprobado que los exámenes que son más difíciles para los humanos también son más difíciles para las máquinas. Finalmente, avanzamos el estado del arte para el conjunto de datos QA4MRE Entrance Exams en varios idiomas.
Palabras clave: Búsqueda de Respuestas; Opción múltiple; Multilingüismo

1 Introduction

Question Answering (QA) has manifold dimensions, depending on the source of the information (i.e. free text vs. knowledge bases), and how the question is to be responded. In this work, we focus on Multiple-Choice QA, where systems has to select the correct answer from a set of candidates according to a given text. This format is usually applied to evaluate language understanding with humans.

To solve this task, the tendency has steered towards deep, attention-based language models, like BERT, or XLNET ((Devlin et al., 2019; Yang et al., 2019)), based on the idea of transformers (Vaswani et al., 2017). These models have boosted all available NLP tasks, becoming the go-to technique in many cases, and QA is not an exception.

These models could be a handicap for underrepresented languages in terms of available resources to train them. In addition, the training of these models requires a huge computation power. Hence, there is a great interest from the research community towards developing multilingual models trained once for many different languages. So far, mono-

ISSN 1135-5948. DOI 10.26342/2020-65-4

lingual models still seem to perform better than multilingual ones for some tasks (Martin et al., 2019; Agerri et al., 2020; Cañete et al., 2020).

In order to use these pre-trained models, systems must be fine-tuned to the target task. Unfortunately, the majority of languages don't have large enough datasets for several tasks, this is the case of Multiple-Choice QA. Given the size of the RACE dataset (Lai et al., 2017), it is possible to fine-tune an English model with it, but you cannot do the same for other languages as the datasets are too small. Such is the case of Spanish or Italian. The creation of these datasets is usually very costly both in time and money. Thus, the majority of collections are either too small to train a deep model or are only available in English (Hsu, Liu, and Lee, 2019).

In this work, we study how to apply the models trained with a dataset in a language to a different collection in another language. For this purpose, we use the RACE collection to train a model and we test the model using the Entrance Exams (EE) datasets, which are available in several languages (Rodrigo et al., 2018).

According to these observations, the objectives of this work are:

i) Compare the results obtained in EE and RACE. In principle, they target similar human language skills: middle and high school English level in the case of RACE and university admission in the case of EE.

ii) Determine if there is a correlation between the difficulty of the exercises for both humans and computers.

iii) Determine if the knowledge learnt by fine-tuning with a collection (RACE in English) can be transferred to perform with another collection (EE English and other languages).

iv) Test the performance of both monolingual and multilingual BERT models once they are trained in one language and evaluated in different ones.

v) Advance the state-of-the-art for the EE task in various languages.

These objectives are motivated by the following research questions:

RQ 1 How monolingual and multilingual models perform for Multiple-Choice QA when they are trained for a specific language to work in a different one?

RQ 2 When monolingual and multilingual models are trained and tested for the same language, is their performance comparable?

RQ 3 Can multilingual models advance the current state-of-the-art for some languages where there is not enough training data?

2 Related Work

Question Answering (QA) is the task of returning a precise and short answer given a Natural Language question. QA can be approached from two main perspectives (Rogers et al., 2020): 1) Open QA, where systems collect evidences and answers across several sources such as Web pages and knowledge bases (Fader, Zettlemoyer, and Etzioni, 2013) and, 2) Reading Comprehension (RC), where the answer is gathered from a single document.

RC systems can be oriented to: (1) extract spans of text with the answer (extractive QA), (2) select an answer from a set of candidates (multiple-choice QA) or (3) generate an answer (generative QA). Extractive QA has received a lot of attention fostered by the availability of popular benchmarks such as SQuAD (Rajpurkar, Jia, and Liang, 2018). On the other hand, generative QA has received less attention given that it is difficult to perform an exact evaluation and there are few datasets (Kočiský et al., 2018).

In this work we focus on Multiple-Choice (MC) QA. Since MC is a common way to measure reading comprehension in humans, the task is very realistic. In fact, the datasets employed in this research (presented ahead) are based on real world exams. Besides, some researches have pointed MC format as a better format to test language understanding of automatic systems (Rogers et al., 2020).

There exists several MC collections, mostly in English. In some cases it involves paying crowd-workers to gather documents and/or pose questions regarding those documents. MCTest (Richardson, Burges, and Renshaw, 2013), for example, proposed for the workers to invent short, children friendly, fictional stories and four questions

with four answers each, including deliberately wrong answers. As a way to encourage a deeper understanding of texts, the QuAIL dataset includes unanswerable questions (Rogers, Kovaleva, and Rumshisky, 2020). Other datasets were created from real world exams. This is the case of the well known MC dataset RACE (Lai et al., 2017), or the multilingual Entrance Exams (Rodrigo et al., 2018), described in more detail in the next Section.

When doing QA, there usually exists the constraint on the language and size of the datasets available. In this sense, many times there is not enough training data to fine-tune a model in a specific language. (Asai et al., 2018) tried to solve this issue by translating the target collection to a language with enough training data and using a QA system trained in the second language. However, this approach relies too much on the quality of the translation.

A common practice to fill this gap is zero-shot learning, which aims to solve a task without receiving any example of that task at the training phase. That is, we fine-tune for task A and evaluate in task B, possibly in another language. Thus, we expect for the knowledge to be transferred from one task to the other (in another language) with minimum overhead.

We have found similar efforts in the literature. Hsu, Liu, and Lee (2019), for example, studies how to train Multi BERT for extractive QA in a language to test it in another language, they obtain promising results. We differ from them in: (1) the languages employed: they test English, Korean and Chinese; (2) the task: they work on extractive QA and; (3) the type of collections: we use collections crafted for human evaluation, which allows us to study how difficulty for humans correlates with difficulty for automatic systems. More specifically, we zero-shot transfer a model from RACE (fine-tune) to Entrance Exams (no training data available) in multiple, unseen languages.

3 Datasets

In our experiments we use RACE and Entrance Exams. Both collections are derived from real human evaluations. The following subsection gives further details of each collection.

3.1 RACE

RACE (Lai et al., 2017) was collected from the English exams for middle (subset named RACE-M) and high (subset named RACE-H) school Chinese students. There are two subsets depending on the level of the exams: *RACE-M* for middle school and *RACE-H* for high school.

The authors proposed it to evaluate the reading comprehension task using MC format. RACE consists of more that 100K questions generated from human experts (English instructors). Table 1 shows the details of RACE. We can see in the table that RACE-H contains more data than RACE-M.

In order to evaluate the difficulty of the collection, the authors employed Amazon Mechanical Turk[1] to annotate question types of a subset. The authors found a higher ratio of reasoning questions with respect to CNN (Hermann et al., 2015), SQUAD and NEWSQA (Trischler et al., 2017), which justifies that RACE is more difficult than those datasets.

3.2 Entrance Exams

The Entrance Exams (EE) data was collected from standardized English examinations for university admission in Japan and used in the Entrance Exams task at CLEF in 2013, 2014 and 2015 (Rodrigo et al., 2018). Exams were created by the Japanese National Center for University Admission Tests. Only the exams with MC format were included in the dataset. We show in Table 2 the number of documents and questions released in each edition, as well as the number for the overall set.

The organizers of the task proposed also the same task in other languages different from English by collecting parallel translations from volunteers at the translation for progress website[2]. EE data is also available in French, Italian, Spanish and Russian. Translations for German are only available for the 2015 dataset. Thus, EE allows to test QA systems in other languages besides English, which is the language where almost all the QA datasets are available. However, EE received only participants for the English and French tasks. Therefore, this paper represents the first attempt to solve EE in the other languages.

[1] https://www.mturk.com/mturk/welcome

[2] `http://www.translationsforprogress.org/`

Dataset	RACE-M			RACE-H			RACE			
Subset	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	All
# documents	6,409	368	362	18,728	1,021	1,045	25,137	1,389	1,407	27,933
# questions	25,421	1,436	1,436	62,445	3,451	3,498	87,866	4,887	4,934	97,687

Table 1: Details of RACE-M, RACE-H and RACE collections

	2013	2014	2015	All
# documents	9	12	19	40
# questions	46	56	89	191

Table 2: Details of Entrance Exams collection

We want to remark also that the EE organizers proposed two kinds of evaluation. The first one is based in traditional QA evaluation, measuring the overall performance of a system over the whole set of questions. The second approach proposed to measure the number of tests passed by a system. According to this approach, each test is made of a document and the questions about it. Then, a system passes a test if it manages to answer correctly to 50% or more of the questions, similar to human evaluations.

4 BERT and Multilingual BERT

BERT and Multilingual BERT (M-BERT from now on) are transformer-based language representation models. They have been pre-trained from unlabeled text to do Masked Language Modeling and Next Sentence Prediction (Devlin et al., 2019). Afterwards, each model can be fine-tuned in specific tasks such as those at Glue (Wang et al., 2018) or QA. Albeit both models share the same architecture (a twelve layer transformer), they were trained with different corpus. BERT was trained with BooksCorpus (800M words) and M-BERT with the Wikipedia in 104 different languages. Even so, both use the same word piece vocabulary (and tokenizer) and have no information about the language in training.

The idea behind M-BERT is to learn all languages at once, delivering a single model capable of operating in multiple languages. There are several caveats with this approach:

1. Languages compete against one another for a fraction of hyper parameters, affecting underrepresented languages. Although oversampling is applied, it is not completely solved (Artetxe, Ruder, and Yogatama, 2019).
2. No language specific knowledge is used to improve in any part of the model. Intuitively, one should be able to pre-train a BERT model in any language, applying language-specific knowledge and improve over M-BERT (e.g.: In (Agerri et al., 2020), the authors employ a basque-specific word piece vocabulary to improve the basque model).

In this paper, we compare both models and their ability to perform zero-shot cross-lingual transfer in multiple-choice QA. We describe the experiments in the next Section.

5 Experiments

To the best of our knowledge, there exists no Spanish MC collections big enough to fine tune a model. So, we have fine-tuned a model on English RACE and test it on other languages.

In our experiment, we use a simple BERT model and a M-BERT. Each model has been fine-tuned over RACE for three epochs, as recommended by the developers. We have employed the well known transformers library from huggingface[3], following the hyper-parameters stated by the first BERT[4] base model result on RACE's leaderboard[5]. Additionally, all the source code is available in a Github repository[6]. Every model was trained on Google Cloud Platform with a Tesla T4 for three epochs.

The experiments followed with each model are:

1. Fine-tune model on RACE train collection, with both high and middle splits.
2. Measure[7] performance on RACE test collection.

[3] https://huggingface.co/transformers/

[4] https://github.com/NoviScl/BERT-RACE

[5] http://www.qizhexie.com/data/RACE_leaderboard.html

[6] https://github.com/m0n0l0c0/race-experiments

[7] We use accuracy

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

Table 3: Accuracy of each model, including baselines: BERT, M-BERT, Random and Longest over each dataset RACE (every split) and Entrance Exams (over every year and language) **German data is a single result, there is data only for 2015

3. Measure performance on EE in: English, Spanish, Italian, French, Russian and German

We have also ran two baselines to establish a lower bound every model should surpass. The first baseline choose every answer at random. Since we have four candidates per question, this baseline achieves an accuracy score of 0.25. The second, following the work of (Rogers et al., 2020), just yields the longest answer.

6 Results

Table 3 shows the results, according to accuracy, obtained with the conducted experiments. We list the results obtained in each dataset (RACE and Entrance Exams) by each employed model (BERT, M-BERT, Random and Longest baselines). In all cases, we report the results for the test split. For Entrance Exams we show the results for all years averaged together.

BERT scores similar in RACE and Entrance Exams, though it obtains its best scores in RACE middle. Even so, it is outperformed by M-BERT in all cases. The latter performs better in RACE than Entrance Exams, which are harder. Both models are above the baselines excluding BERT with Russian EE.

Both models' scores visibly decrease when raising the education grade. RACE middle is the highest scored, which corresponds to middle school exams, the easiest of the three collections for humans. Following, RACE high and, in the last place, Entrance Exams, which are tests for university entrance. This tendency matches that of the humans, when increasing in difficulty, students score lower.

Both models are above the baselines, excluding BERT's Entrance Exams - Russian result which is the worst by far.

Among the Entrance Exams collection, the English version obtains the highest score for both models. Taking into account that both models were fine-tuned with a single task in English, this was expected. On the other hand, EE Russian was the lowest scored collection. Our intuition is that it is due to using a different alphabet.

No model had previous clue of Entrance Exams task, it was zero-shot transferred from RACE. That is only available in English.

This means that in the case of BERT, there are languages that it has never seen before, because it is monolingual. Furthermore, it's performance worsens hardly when exposed to unseen languages, specially with Russian and German, this is the expected behavior since languages with very different semantics are not tokenized nor understood correctly by the model (Artetxe, Ruder, and Yogatama, 2019; Agerri et al., 2020). This is the case of Russian.

M-BERT scores above 0.4 in all cases but Russian. Additionally, it is very close to passing the exam (to answer correctly at least 50% of the questions) for English.

Tables 4, 5 and 6 show the results of EE divided by years (from 2013 to 2015). We give results according to accuracy and the proportion of tests (a document with their corresponding questions) passed (an accuracy of at least 0.5). These results correspond to the two evaluation perspectives applied in EE and described in Section 3.2. We include results in each language for each model, the two baselines and the best performing system at EE (in each edition). There are results from

Entrance Exams 2013

	BERT		MultiBERT		Random		Longest		NIIJ-3*	
	Acc	Tests	Acc	Tests	Acc	Tests	Acc	Tests	Acc	Tests
English	**0.43**	0.22	0.41	0.44	0.25	0	0.22	0.00	0.35	0.33
Spanish	0.37	0.33	**0.43**	0.22	0.25	0	0.22	0.11	-	-
Italian	0.28	0.11	**0.33**	0.22	0.25	0	0.28	0.22	-	-
French	0.35	0.11	**0.43**	0.33	0.25	0	0.17	0.00	-	-
Russian	0.22	0.11	**0.26**	0.00	0.25	0	0.17	0.00	-	-

Table 4: Accuracy of each model and proportion of passed tests, 2013 edition had 9 tests in total. Results from all models and baselines: BERT, M-BERT, Random and Longest over every language in Entrance Exams 2013. The best previous work (*) on (Rodrigo et al., 2018) from the National Institute of Informatics of Japan (Li et al., 2013). They originally presented results only for English

Entrance Exams 2014

	BERT		MultiBERT		Random		Longest		Synapse*	
	Acc	Tests	Acc	Tests	Acc	Tests	Acc	Tests	Acc	Tests
English	0.50	0.58	**0.52**	0.67	0.25	0	0.30	0.33	0.45	0.58
Spanish	0.38	0.33	**0.45**	0.50	0.25	0	0.34	0.25	-	-
Italian	0.32	0.33	**0.43**	0.33	0.25	0	0.29	0.17	-	-
French	0.30	0.17	0.48	0.50	0.25	0	0.30	0.25	**0.59**	0.75
Russian	0.30	0.17	0.32	0.17	0.25	0	**0.34**	0.17	-	-

Table 5: Accuracy of each model and proportion of passed tests, 2014 edition had 12 tests in total. Results from all models and baselines: BERT, M-BERT, Random and Longest over every language in Entrance Exams 2014. The best previous work (*) on (Rodrigo et al., 2018) from Synapse (Laurent et al., 2014). They originally presented results only for French and English

Entrance Exams 2015

	BERT		MultiBERT		Random		Longest		Synapse*	
	Acc	Tests	Acc	Tests	Acc	Tests	Acc	Tests	Acc	Tests
English	0.52	0.63	0.53	0.63	0.25	0	0.19	0.21	**0.58**	0.84
Spanish	0.36	0.32	**0.46**	0.53	0.25	0	0.30	0.26	-	-
Italian	0.27	0.26	**0.48**	0.53	0.25	0	0.27	0.16	-	-
French	0.28	0.11	0.40	0.47	0.25	0	0.27	0.32	**0.56**	0.84
Russian	0.26	0.11	**0.39**	0.47	0.25	0	0.28	0.32	-	-
German**	0.37	0.42	**0.45**	0.58	0.25	0	0.26	0.16	-	-

Table 6: Accuracy of each model and proportion of passed tests, 2015 edition had 19 tests in total. Results from all models and baselines: BERT, M-BERT, Random and Longest over every language in Entrance Exams 2015. The best previous work (*) on (Rodrigo et al., 2018) from Synapse (Laurent et al., 2015). They originally presented results only for French and English
**This is the only year with German data

previous systems for English in all editions and French in 2014 and 2015. Thus, our work is setting the results for these collections in several languages.

BERT model shows its best result in English, 2013, but M-BERT passes a higher proportion of tests (almost twice). This means that BERT finds the correct answer for more questions than M-BERT but distributed across just a few documents, obtaining higher grades. However, M-BERT scores better in general, passing more than 44% of the tests.

The rest of results for 2013 are dominated by M-BERT model, which states a new best result in English, outperforming the previous systems. In the case of the second campaign, M-BERT lands second on the results table for French (where the best result is obtained by (Laurent et al., 2014)), and surpasses previ-

ous results in English. M-BERT model also sets the new results in Spanish and Italian (the best result for Russian is achieved by the longest baseline). From table 6, the best results in English and French go for Synapse (Laurent et al., 2015), who developed a complex system including background knowledge and trained over the previous dataset of EE. Thus, it seems that simple pre-trained models, transferred over a different dataset, are close to those results but not enough. For the rest of languages, the best results come from M-BERT model, that sets the state-of-the-art without requiring a dataset in those languages.

Overall, we can observe a best performance of M-BERT with respect to BERT, which is focused on the English language. In fact, according to our experiments, we can fine-tune M-BERT in English for MC using one dataset and transfer that knowledge to datasets in other languages.

7 Conclusions and Future Work

The results obtained show that both monolingual and multilingual models can be fine-tuned for task and transferred to another task in the same language. Furthermore, multilingual models are transferable also to different languages.

Also, we obtain evidence that systems performance is hampered by exams difficulty in the same way human grades do.

In this work we established the state-of-the-art results over the Entrance Exams task in four more languages.

We would like to continue pursuing methods to cope with low resource languages. To do so, we will continue exploring how fine-tuned transformer bodies can be transferred to reuse knowledge about specific tasks, following the lead of (Artetxe, Ruder, and Yogatama, 2019; Otegi et al., 2020).

Acknowledgments

This work has been funded by the Spanish Research Agency under CHIST-ERA LIHLITH project (PCIN-2017-085/AEI) and deepReading (RTI2018-096846-B-C21 / MCIU/AEI/FEDER,UE).

References

Agerri, R., I. S. Vicente, J. A. Campos, A. Barrena, X. Saralegi, A. Soroa, and E. Agirre. 2020. Give your Text Representation Models some Love: the Case for Basque. mar.

Artetxe, M., S. Ruder, and D. Yogatama. 2019. On the Cross-lingual Transferability of Monolingual Representations. oct.

Asai, A., A. Eriguchi, K. Hashimoto, and Y. Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *CoRR*, abs/1809.03275.

Cañete, J., G. Chaperon, R. Fuentes, and J. Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *to appear in PML4DC at ICLR 2020*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, oct. Association for Computational Linguistics.

Fader, A., L. Zettlemoyer, and O. Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1608–1618, Sofia, Bulgaria, August. Association for Computational Linguistics.

Hermann, K. M., T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Hsu, T.-Y., C.-L. Liu, and H.-y. Lee. 2019. Zero-shot Reading Comprehension by Cross-lingual Transfer Learning with Multi-lingual Language Representation Model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5933–5940, Hong Kong, China. Association for Computational Linguistics.

Kočiský, T., J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, and E. Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Lai, G., Q. Xie, H. Liu, Y. Yang, and E. Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 785–794, apr.

Laurent, D., B. Chardon, S. Nègre, C. Pradel, and P. Séguéla. 2015. Reading comprehension at entrance exams 2015. In L. Cappellato, N. Ferro, G. J. F. Jones, and E. SanJuan, editors, *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*, volume 1391 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Laurent, D., B. Chardon, S. Nègre, and P. Séguéla. 2014. French Run of Synapse Développement at Entrance Exams 2014. In *CLEF (Working Notes)*, pages 1415–1426.

Li, X., R. Tian, N. L. T. Nguyen, Y. Miyao, and A. Aizawa. 2013. Question Answering System for Entrance Exams in QA4MRE. In *CLEF (Working Notes)*. Citeseer.

Martin, L., B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot. 2019. CamemBERT: a Tasty French Language Model. nov.

Otegi, A., A. Agirre, J. A. Campos, A. Soroa, and E. Agirre. 2020. Conversational Question Answering in Low Resource Scenarios: A Dataset and Case Study for Basque. In *12th International Conference on Language Resources and Evaluation*.

Rajpurkar, P., R. Jia, and P. Liang. 2018. Know What You Don{'}t Know: Unanswerable Questions for {SQ}u{AD}. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Richardson, M., C. J. C. Burges, and E. Renshaw. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. Technical report, nov.

Rodrigo, A., A. Peñas, Y. Miyao, and N. Kando. 2018. Do systems pass university entrance exams? *Information Processing & Management*, 54(4):564–575, jul.

Rogers, A., O. Kovaleva, M. Downey, and A. Rumshisky. 2020. Getting closer to ai complete question answering: A set of prerequisite real tasks.

Rogers, A., O. Kovaleva, and A. Rumshisky. 2020. A Primer in BERTology: What we know about how BERT works. feb.

Trischler, A., T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada, August. Association for Computational Linguistics.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pages 5998–6008.

Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 {EMNLP} Workshop {B}lackbox{NLP}: Analyzing and Interpreting Neural Networks for {NLP}*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., jun, pages 5754–5764.