Supervised Domain Adaptation for Extractive Question Answering in Spanish

Santiago Máximo[1]

[1]*Universidad de la República, Montevideo, Uruguay*

Abstract

Recent releases of pre-trained language models and Question Answering (QA) datasets have led to rapid improvements in Extractive QA. This paper describes the work done for QuALES, part of IberLEF 2022, a task to automatically find answers to questions in Spanish from news text related to Covid-19. We present an approach mainly centered on transfer learning applied to BETO and RoBERTa-base-bne based models. The models were fine tuned on different combinations of Spanish QA datasets. Our submission achieved third place in QuALES challenge for Exact Match and F1-Score metrics.

Keywords

Question Answering, Domain Adaptation, Transformers, SQuAD

1. Introduction

Question Answering (QA), as well as other traditional Natural Language Processing tasks, has seen considerable progress in recent times due to the appearance of Transformers.

QuALES Competition [1] focuses on Extractive Question Answering, the task of extracting a span of text from a context as the answer to a question.

Although most of the research is related to the English language, in the last few years not only Spanish language models but also QA Spanish datasets have been developed, which leaves the door open to new advances in the field.

Given the difficulty of creating large domain-specific QA datasets, this paper explores the effectiveness of using general-purpose QA datasets, together with small specialized datasets, when fine tuning models for restricted domains. Particularly, QuALES challenge was evaluated on a corpus of news in Spanish related to the Covid-19 domain.

2. Methodology

2.1. Datasets

In addition to the train dataset provided by the organizers of the challenge, we used three other available Spanish QA datasets for training our models.

SQuADesV2: an automatic translation of the Stanford Question Answering Dataset into Spanish [2]. The original English SQUAD v1.1 [3] is a reading comprehension dataset consisting

IberLEF 2022, September 2022, A Coruña, Spain.

santiagomaximoc@gmail.com (S. Máximo)

of 100,000 questions posed by crowdworkers on a set of Wikipedia articles. SQUAD v2 [4] combines the 100,000 questions in SQuAD v1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. SQuAD v2 Spanish translated version was used in this work.

SQAC: an extractive QA dataset with only answerable questions. It was created from texts extracted from the Spanish Wikipedia, encyclopedic articles, newswire articles from Wikinews, and the Spanish section of the AnCora corpus. It consists of 18,817 questions with the annotation of their answer spans from 6,247 textual contexts, following the guidelines from SQuAD v1.1 [5].

NewsQAes: a Spanish translation of the NewsQA dataset [6]. The original English NewsQA [7] is a reading comprehension dataset of 120,000 QA pairs from CNN news articles. The Spanish NewsQA used in this work does not include unanswerable questions.

CovidQA: the QuALES training set consisting of about 1,000 questions from news articles in Spanish related to the Covid-19 domain, including unanswerable questions.

2.2. Models

For our experiments, we selected two base Spanish Language Models: Beto and RoBERTa-base-bne.

BETO: is a masked language model trained on a big Spanish corpus, including the Spanish portion of ParaCrawl, EUBookshop and Wikipedia [8]. BETO was built upon ideas from the original BERT architecture [9], Bidirectional Encoder Representations from Transformers, and also incorporated techniques that have been successful in RoBERTa [10], in particular dynamic masking.

RoBERTa-base-bne: is a transformer-based masked language model based on the RoBERTa architecture and has been pre-trained using a large Spanish corpus compiled from the web crawlings performed by the National Library of Spain (Biblioteca Nacional de España) from 2009 to 2019 [5].

Not only the base versions of Beto and RoBERTa-base-bne were selected, but also other available fine-tuned variants of these models trained on the SQAC and SQuADesV2 datasets. Table 2.2 shows all the pretrained models taken from the Hugging Face Hub [11] .

2.3. Preprocessing

The questions, answers and contexts were tokenized using the corresponding model Tokenizer.

However, there are a few preprocessing steps that are particular to the Question Answering task. The dataset may include contexts that exceed the maximum input length of the model. If the context is truncated, the answer may be outside of the truncated text. To fix this problem, the context has to be split into chunks. Moreover, in order not to split context in the middle of the answer, an overlap between chunks needs to be included. The guidelines for the preprocessing implementation were followed from the Hugging Face Course [12].

In the case of the NewsQAes, some other modifications were applied to the contexts. Opening sentences containing redundant location information were removed. Besides, occurrences of the

Hugging Face Model	Base Model	Fine-Tunings	
		First	Second
PlanTL-GOB-ES/roberta-base-bne	RoBERTa-base-bne		
PlanTL-GOB-ES/roberta-base-bne-sqac	RoBERTa-base-bne	SQAC	
hackathon-pln-es/roberta-base-bne-squad2-es	RoBERTa-base-bne	SQuADesV2	
dccuchile/bert-base-spanish-wwm-cased	BETO		
IIC/beto-base-spanish-sqac	BETO	SQAC	
MMG/bert-base-spanish-wwm-cased-finetuned-sqac-finetuned-squad2-es	BETO	SQAC	SQuADesV2
MMG/bert-base-spanish-wwm-cased-finetuned-squad2-es	BETO	SQuADesV2	
MMG/bert-base-spanish-wwm-cased-finetuned-spa-squad2-es-finetuned-sqac	BETO	SQuADesV2	SQAC

Table 1
Description of Hugging Face models used in this work.

word "CNN" between brackets were also deleted. Finally, the NewsQAes Dataset was converted to SQuAD format in order to apply the same techniques as to the other datasets.

2.4. Postprocessing

During Preprocessing one context may have been split into several chunks. Once all possible answers for that context have been scored, the one with the best score is selected. The guidelines for the Postprocessing implementation were followed from the Hugging Face Course [12].

2.5. Training

We conducted the experiments with consecutive fine tunings, consisting of different combinations of the available QA datasets. When included in an experiment, the CovidQA dataset was located in the last step of the chain.

The models were implemented using the Transformers Library [13], with document stride set to 128, maximum input length 384, learning_rate 3e-5, and AdamW optimizer.

Models trained on SQuADesV2 and NewsQAes, were trained for 2 epochs with batch size of 24.

Models trained on CovidQA, were trained for 5 epochs with batch size of 16.

It was not necessary to perform any training on the SQAC dataset, as there were already pre-trained models available in the Hugging Face Hub.

Base Model	Fine-Tunings				Scores	
	First	Second	Third	Fourth	EM	F1
RoBERTa-base-bne	SQAC	SQuADesV2	CovidQA		50.58	62.33
BETO	SQAC	SQuADesV2	CovidQA		50.45	60.98
BETO	SQAC	CovidQA			48.51	57.74
BETO	SQuADesV2	CovidQA			47.99	58.46
BETO	SQAC	NewsQAes	SQuADesV2	CovidQA	47.87	57.59
BETO	NewsQAes	SQuADesV2	CovidQA		47.87	57.23
RoBERTa-base-bne	SQuADesV2	CovidQA			47.09	55.96
BETO	SQuADesV2	SQAC	CovidQA		46.57	54.92
RoBERTa-base-bne	SQAC	CovidQA			42.04	51.21
RoBERTa-base-bne	SQAC				41.14	49.19
BETO	SQuADesV2	SQAC			40.75	51.38
RoBERTa-base-bne	SQuADesV2				40.36	52.83
BETO	SQAC	NewsQAes	SQuADesV2		37.52	48.33
BETO	SQAC				36.22	45.16
BETO	SQAC	SQuADesV2			35.58	44.22
RoBERTa-base-bne	SQAC	SQuADesV2			35.19	43.26
BETO	NewsQAes	SQuADesV2			33.89	42.76
BETO	SQuADesV2				31.69	38.39
BETO	NewsQAes				21.99	33.87

Table 2
Statistics over the development set.

	EM	F1
Average with CovidQA	47.66	57.38
Average without CovidQA	35.43	44.94

Table 3
Performance comparison between models trained with and without CovidQA dataset.

3. Results

The results reported in tables 3, 3, and 3 were obtained after testing our models on the development set with gold labels published in the Codalab's competition page. We used the F1-Score and Exact Match metrics provided by the Transformers Library [13] for SQuAD V2.

In all scenarios where CovidQA dataset was included in the chain of fine-tunings, the performance of the model improved, somewhat expected due to its similarity to the evaluation dataset. As it is shown in table 3, on average, Exact Match improved by 12.23% while F1-Score increased by 12.44%.

The results obtained after training the base models solely on CovidQA Dataset are presented in table 3, which indicates that the use of general purpose datasets improve the robustness of the models.

Base Model	EM	F1
RoBERTa-base-bne	37,26	41,92
BETO	32.47	40.15

Table 4
Base models trained solely on CovidQA Dataset.

Team Model Ranking	Answerable Questions EM	Answerable Questions F1	Unanswerable Questions EM	Unanswerable Questions F1
#1	40.65	57.60	87.02	87.02
#2	42.52	56.20	86.26	86.26

Table 5
Performance comparison between answerable and unanswerable questions, on the development set.

Regarding the NewsQAes dataset, unstable results were obtained, with performance improving in some experiments and dropping in others. Although its source text comes from news, similar to the CovidQA evaluation dataset, it has some shortcomings as it is a translation and does not include unanswerable questions.

The two best models were obtained as a result of fine-tuning on the same chain of datasets. This winning combination includes the SQAC dataset at the beginning, followed by SQuADesV2, and ending with CovidQA. Exchanging the positions of the SQAC and SQuADesV2 results in a performance degradation. This behavior can be explained due to the fact that the SQuADesV2 contains unanswerable questions, and placing it close to the end of the chain may prevent the model from losing the ability to return empty answers.

The presence of the SQAC dataset at first in the top combinations is another interesting finding. Although SQAC does not contain unanswerable questions, its source text is written originally in Spanish, resulting in a high-quality dataset. Therefore, by placing it at the beginning of the chain, it may help the model not only to learn how to answer questions, but also to better handle native Spanish.

In order to evaluate the behavior of the models against answerable and unanswerable questions, we developed some customized scripts. The resulting metrics for the top two solutions evaluated on the development set, reported in table 3, suggests that our models respond more accurately to unanswerable questions.

Finally, table 3 shows the results obtained from the leaderboard of the competition for the evaluation phase, which were consistent with those obtained for the development set. Our model achieved third place in both metrics: Exact Match and F1-Score.

4. Conclusion and Future Work

In this paper, we present performance differences between Spanish pre-trained language models fine-tuned on various question answering datasets, and evaluated for the task of automatically finding answers to questions in Spanish from news text related to Covid-19.

User	EM
sebastianvolti	53.49
ichramm	46.77
smaximo	**45.98**
Bernardo	44.27
avacaondata	39.92

User	F1
sebastianvolti	72.82
Bernardo	61.59
smaximo	**61.42**
rigoberta	58.77
avacaondata	58.73

Table 6
QuALES results for the evaluation phase.

Our RoBERTa-base-bne fine-tuned consecutively on SQAC, SQuADesV2, and CovidQA datasets, obtained the third best performance in the QuALES Challenge for Exact Match and F1-Score metrics.

Our experiments reveal that despite the small amount of in-domain training data, the use of these sources in conjunction with general-domain QA datasets can significantly improve the performance of the models, reducing the gap between human evaluation results. However, considerably more work will need to be done in order to implement robust language models for domain-specific Extractive Question Answering.

Unsupervised domain adaptation, could be an area of future work, taking advantage, for example, of the large amount of Spanish news available.

In addition, based on the winning chain of datasets, an exhaustive hyperparameter search can be executed in order to improve the effectiveness of the models. A joint multi-dataset training could be another future step.

Finally, more focus on generating high quality Spanish QA datasets, containing unanswerable questions, might prove beneficial.

References

[1] A. Rosá, L. Chiruzzo, L. Bouza, A. Dragonetti, S. Castro, M. Etcheverry, S. Góngora, S. Goycoechea, J. Machado, G. Moncecchi, J. J. Prada, D. Wonsever, Overview of QuALES at IberLEF 2022: Question Answering Learning from Examples in Spanish, Procesamiento del Lenguaje Natural 69 (2022).

[2] C. Casimiro Pio, C.-j. Marta R., F. Jose A. R., Automatic Spanish Translation of the SQuAD Dataset for Multilingual Question Answering, arXiv e-prints (2019) arXiv:1912.05200v1. arXiv:1912.05200v2.

[3] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, arXiv e-prints (2016) arXiv:1606.05250. arXiv:1606.05250.

[4] P. Rajpurkar, R. Jia, P. Liang, Know what you don't know: Unanswerable questions for squad, CoRR abs/1806.03822 (2018). URL: http://arxiv.org/abs/1806.03822. arXiv:1806.03822.

[5] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas,

Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022) 39–60. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405.

[6] newsqa-es, Code to rebuild the newsqa-es dataset: a spanish version of the newsqa dataset, 2022. URL: https://github.com/pln-fing-udelar/newsqa-es.

[7] A. Trischler, T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, K. Suleman, Newsqa: A machine comprehension dataset, in: Proceedings of the 2nd Workshop on Representation Learning for NLP, 2017, pp. 191–200.

[8] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[9] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[11] Huggingface hub, 2020. URL: https://huggingface.co/models.

[12] Fast tokenizers in the qa pipeline, 2021. URL: https://huggingface.co/course/chapter6/3b?fw=pt.

[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.