Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 273-280 recibido 05-07-2022 revisado 21-07-2022 aceptado 24-07-2022

Overview of QuALES at IberLEF 2022: Question Answering Learning from Examples in Spanish

Overview de QuALES en IberLEF 2022: Preguntas y Respuestas Automáticas sobre Ejemplos en Español

Aiala Rosá[1], Luis Chiruzzo[1], Lucía Bouza[1], Alina Dragonetti[1], Santiago Castro[2], Mathias Etcheverry[1], Santiago Góngora[1], Santiago Goycoechea[1], Juan Machado[1], Guillermo Moncecchi[1], Juan José Prada[1], Dina Wonsever[1]
[1]Universidad de la República, Montevideo, Uruguay
{aialar, luischir, lucia.bouza, alina.dragonetti, mathiase, sgongora, sgoycoechea, juan.machado, gmonce, prada, wonsever}@fing.edu.uy
[2]University of Michigan, Anne Arbor, USA
sacastro@umich.edu

Abstract: We present the results of the QuALES task, which addresses the problem of Extractive Question Answering from texts. For both training and evaluation we use the QuALES corpus, a corpus of Uruguayan media news about the Covid-19 pandemic and related topics. We describe the systems developed by seven participants, all of them based on different BERT-like language models. The best results were obtained using the multilingual RoBERTa model pre-trained with SQUAD-Es-V2, with a fine tuning on the QuALES corpus.
Keywords: Question Answering for Spanish, Language Models, Datasets for Question Answering.

Resumen: Presentamos los resultados de la tarea QuALES, que aborda el problema de Búsqueda de Respuestas extractiva a partir de textos. Tanto para entrenamiento como para evaluación utilizamos el corpus QuALES, un corpus de noticias de medios uruguayos sobre la pandemia por Covid-19 y temas relacionados. Describimos los sistemas desarrollados por siete participantes, todos ellos basados en diferentes modelos de lenguaje tipo BERT. Los mejores resultados se obtuvieron usando el modelo RoBERTa multilingüe preentrenado con SQUAD-Es-V2, con una fine tuning sobre el corpus QuALES.
Palabras clave: Búsqueda de Respuestas en Español, Modelos de Lenguaje, Corpus para Búsqueda de Respuestas.

1 Introduction

Question Answering (QA) is a classical Natural Language Processing task that is currently gaining great relevance. QA can be roughly divided into two main categories (Jurafsky and Martin, 2021): semantic analysis, where the question is transformed to a query to a knowledge database; and open domain question answering, where, starting from a question written in natural language and a set of documents, the answer to the question is obtained using information retrieval and information extraction techniques.

Open domain question answering involves two main stages: a) getting the relevant documents, generally using methods from the Information Retrieval field (IR) (Manning, Raghavan, and Schütze, 2010), possibly one of the most widely studied topics in NLP, with web search engines as their most noticeable product, b) extracting the answer from those documents. Each of these stages has its own challenges, and the whole task requires a successful outcome for each of them and for their integration.

In this task we address the problem of extractive QA in Spanish, based on a corpus of a specific domain: press news about the Covid-19 pandemic. We focus on the second stage of the task: given a text, extracting the

ISSN 1135-5948. DOI 10.26342/2022-69-24

answer to a question, if there is one.

The rest of the paper is structured as follows: section 2 describes the background of QA, focusing on QA for Spanish; section 3 describes the corpus created for this task and some other resources; section 4 describes the QuALES task; section 5 presents the participants systems and analyzes the results; and, finally, section 6 shows some conclusions.

2 Background

Starting last decade, along with the popularization of distributional semantic methods based on neural networks (Le and Mikolov, 2014; LeCun, Bengio, and Hinton, 2015), this kinds of methods started to be applied to the QA task, achieving significant results improvement (Yu et al., 2014; Min et al., 2018; Xiong, Zhong, and Socher, 2017; Seo et al., 2016).

All these supervised learning approaches were possible due to the existence of research oriented publicly available datasets. These datasets have enabled not only model training, but also constant monitoring of this area's state of the art. Probably the most popular is SQuAD (Rajpurkar et al., 2016). To build this dataset, annotators were presented with a Wikipedia paragraph and asked to write questions that could be answered from the given text. Natural Questions (Kwiatkowski et al., 2019b) was created from actual Google Search queries, where annotators marked the answer into Wikipedia article snippets. TriviaQA (Joshi et al., 2017) contains a set of Trivia questions and answers. CuratedTREC (Baudiš and Šedivý, 2015) dataset generated by the QA track of the NIST TREC conferences contains questions and answers. NewsQA (Trischler et al., 2016) is a machine comprehension dataset of over 100,000 human-generated question-answer pairs, based on set of over 10,000 news articles from CNN.

In the last few years, after the publication of models based on the Transformers architecture (Vaswani et al., 2017) for solving sequence to sequence transformation problems, and particularly language models such as BERT (Devlin et al., 2018) and ALBERT (Lan et al., 2019), there has been a new push in system performance, particularly for the English language. These kinds of models are trained in an self-supervised way, using large volumes of data and computing power. After that stage (called pretraining), they can be easily fine-tuned to apply them to different tasks. Regarding this shared task, we are particularly interested in fine-tuning them to the open domain question answering task.

The study of the QA area is currently very active, as evidenced by the inclusion of a tutorial[1] on this topic in ACL 2020, the main NLP event worldwide.

Throughout the last few years, several QA related tasks have been proposed. Since 2015, one of the tasks of each SemEval annual international workshop on Semantic Evaluation has been related to some form of the QA Task. For example, SemEval-2015 Task 3: "Answer Selection in Community Question Answering" (Nakov et al., 2019b) proposed, given a question, to classify a certain answer as good, bad, or potential, and answer yes/no questions. The challenge was proposed for Arabic and English. SemEval-2017 Task 3: "Community Question Answering" (Nakov et al., 2019a) proposed three different subtasks: Question-Comment Similarity, Question-Question Similarity, and Question-External Comment Similarity. Additionally, for the Arabic language, another task was added: reranking correct answers for a new question.

SemEval-2022 includes the task "Competence-based Multimodal Question Answering" (Task 09)[2], designed to query how well a system understands the semantics of recipes derived from the R2VQ dataset, a multimodal dataset of cooking recipes and videos.

In (Reddy, Chen, and Manning, 2019) the CoQA (Conversational Question Answering) dataset is presented as a challenge. The dataset includes 127k questions with answers, obtained from 8k conversations about text passages. The goal of the CoQA challenge is to measure the ability of machines to understand a text passage and answer a series of interconnected questions that appear in a conversation.

The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) website includes the highest performing systems on the dataset, measuring Exact Match and

[1] https://github.com/danqi/acl2020-openqa-tutorial

[2] https://competitions.codalab.org/competitions/34056

F1 values (see the next section for a description of these metrics). These systems should answer reading comprehension questions, including questions that do not have an answer on the dataset.

Based on Google's Natural Questions Dataset (Kwiatkowski et al., 2019a), the Tensorflow 2.0 machine learning platform includes a Question Answering competition, where the goal is to predict short and long answer responses to real questions about Wikipedia articles. Using the same dataset, the 2020 NeurIPS Conference an open domain question answering challenge (Min et al., 2021) was also proposed, including three tracks where the objective is to build self-contained question answering systems.

For English, the BioASQ challenge for 2022 proposes a task relative to the Covid-19 domain, using a dataset composed by biomedical articles.

QA research for Spanish has evolved much more slowly. However, similar language resources have been created for this language, which makes us think it is possible to study and fine-tune current architectures to obtain competitive results. In particular, there is a recently developed version of BERT for Spanish, dubbed BETO (Cañete et al., 2020), and a version of SQuAD (the main dataset for training and evaluating open domain QA systems) translated to Spanish (Rajpurkar et al., 2016; Carrino, Costa-jussà, and Fonollosa, 2019). The Spanish Question Answering Corpus (SQAC) is an extractive QA dataset created from texts extracted from a mix from different news-wire and literature sources, and it includes 18,817 questions with the annotation of their answer spans from 6,247 textual contexts (Gutiérrez Fandiño et al., 2022).

From 2003 to 2014, the CLEF Question Answering Track has proposed different campaigns related to question answering, some of which included Spanish datasets. For example, together with the CLEF 2009 forum, ResPubliQA, a Question Answering Task over European legislation was proposed (Peñas et al., 2009). The task consisted of extracting a relevant paragraph of text that included the answer to a natural language question. During CLEF 2010, the task was expanded (Peñas et al., 2010) to include an answer selection task (i.e. besides retrieving the relevant paragraph, systems were required to identify the exact answer to the question). It also proposed several cross-lingual tasks, working on two multilingual parallel corpus: the JRC-ACQUIS Multilingual Parallel Corpus (10,700 parallel and aligned documents), and the Europarl collection (150 parallel and aligned document per language), with 200 question-answer pairs provided for evaluation.

Unlike the task we present here, the CLEF tasks have addressed domain-general questions, or questions for some specific domains, but different from the one selected for QuALES. In addition, they have worked with smaller amounts of training and testing data. Some of these CLEF tasks have some characteristics that differ from our proposal, such as datasets oriented to answer multiple choice questions, or natural language questions to be answered from DBPedia structured data (instead of plain text), among other.

3 Corpus

We provided a corpus of around 2,600 question-answer pairs (the QuALES corpus). The training set contains 1,000 of these pairs, while the dev and test sets have around 800 pairs each. Participants could use any other data for training as well, in particular SQuAD (Rajpurkar et al., 2016) or NewsQA (Trischler et al., 2016). The data is available at the Codalab competition site[3].

The QuALES corpus is original and it was created manually by the members of the team and students. It is a Question-Answering corpus in Spanish obtained from a set of Covid-19 related news published in two important news media from Uruguay (La Diaria[4] and Montevideo Portal[5]). It consists of a set of factoid questions mostly about Covid-19 and its repercussions in Uruguay and the world. Table 1 shows the statistics of the dataset.

The corpus annotation was made in two stages: first, we annotated questions by reading only the title and first sentence of the article; then, we thought of questions derived from the reading of the whole article. For each question, we annotated the answer found (if there was any) and the whole sentence context which included it. For the annotation of the answer, we selected the shor-

[3]https://codalab.lisn.upsaclay.fr/competitions/2619

[4]https://ladiaria.com.uy/

[5]https://www.montevideo.com.uy/

Split	Train	Dev	Test
Articles	176	146	143
Questions	948	773	759
Answers	1000	800	821
Empty answers	165	132	103

Table 1: Statistics of the QuALES corpus showing number of articles, number of questions, total number of answers and total number of questions without answers (empty answers) by split.

test span of text contained in the sentence that consisted in a complete answer for the question. All the answers were directly extracted from the text. Some questions may have more than one answer in a given text, in such cases, a set of answers is generated for this question.

We measured inter-annotator agreement between six pairs of annotators. Each pair answered a set of 25 questions, generating a total of 150 questions with two different annotator answers. We obtained an average Exact Match of 0.61 and an average F1 of 0.76. These results are quite low, which shows the complexity of the task, even for humans. The difference between Exact Match and F1 shows the difficulty in defining the limits of the answer, in general the differences are due to the inclusion or not of elements such as prepositions or determiners. The low F1 shows that selecting the fragment that contains the answer, or deciding that a certain fragment has no answer in the text, is also a highly complex task.

We also published some resources to automatically generate a Spanish version of the NewsQA (Trischler et al., 2016) corpus. The complete NewsQA corpus was translated using a machine translation model and after that we aligned the answers. This alignment stage is necessary because, when translating each fragment with its associated question and answer, the substring corresponding to the answer within the fragment, can be translated differently from the associated answer, which is translated decontextualized. In our translation of the corpus, this alignment problem was detected in 49 % of the cases. To solve this problem we worked on two approaches: on the one hand we trained a neural model from pairs of aligned texts, and, on the other hand, we tested some heuristics defined from the analysis of different examples. In order to evaluate the two approaches, we performed a manual evaluation of a subset of 2,000 question-answer pairs. A portion of this curated corpus was used for parameter tuning of the neural model. The neural model for alignment achieved better results than the heuristics approach. Due to licensing issues, it is not possible to provide a link to this dataset, but the resources to recreate this process are available at our github repository[6].

4 Task

The aim of the QuALES task is to develop question answering systems that can answer questions based on news articles written in Spanish. The systems get a full news article and a question, and must find the shortest span of text in the article (if it exists) that answers the question. It should be noted that for some questions there may not be an answer in the given text. está hablando. The training, development and test datasets were generated from the QuALES corpus, as mentioned above. Originally, we planned to have two separate corpora for evaluation, but seeing that the texts often contain Covid-19 related news mixed with other topics, we decided to annotate only one set. Most of the questions in the dataset are about Covid-19 matters, but some of them are also about other topics.

Table 2 shows a sample text with two questions. The answer to one of the questions can be found in the text, while the other is not present.

As one of our evaluation metrics, we measure average Exact Match for all the dataset instances, following the approach of SQuAD (Rajpurkar et al., 2016). We also report, following (Reddy, Chen, and Manning, 2019), the macro-average F1 score of word overlap: we compare each individual prediction against the different human gold standard answers and select the maximum value as system F1 score for that instance; the system performance is the macro-average of all those F1 scores. Some determiners, specifically, definite articles, and punctuation marks were ignored when calculating this evaluation metric.

Some of the questions in the dataset have more than one possible answer, but the systems are expected to generate at most only

[6]https://github.com/pln-fing-udelar/newsqa-es

Comenzaron las clases presenciales en 344 escuelas rurales, con baja asistencia. A las 8.45 dos perros paseaban por el patio de la escuela rural 27 de La Macana, en Florida. Dos maestras con túnicas blancas y tapabocas esperaban a los alumnos que reanudarían las clases presenciales luego de cinco semanas de conexión virtual. Ya estaba instalado el micrófono y el parlante en el patio, habían llegado los inspectores regionales junto con la directora general del Consejo de Educación Inicial y Primaria (CEIP), Irupé Buzzetti, que junto a la prensa local esperaban a los niños. De los 28 alumnos que asisten regularmente, 14 habían dicho que no iban a ir y los otros no habían confirmado. A las 9.00, cuando debían comenzar las clases en la escuela de La Macana, no había ningún niño. (...) La situación de La Macana se repitió en varias de las escuelas que abrieron este miércoles. De las 547 escuelas habilitadas abrieron 344, confirmó a la diaria Limber Santos, director del departamento de Educación Rural del CEIP. De esas escuelas, cerca de 90 no recibieron alumnos; Santos estimó que en la mañana del miércoles 1.030 niños concurrieron a las escuelas, de un total de 3.900 que concurren a las 547 habilitadas y de 2.838 alumnos que tienen matriculadas las 344 escuelas que abrieron. La asistencia, por tanto, llegó a 36 % en el primer día.
Q1: *¿Cuántas escuelas rurales hay en Uruguay?* A1: *De las [547] escuelas habilitadas abrieron 344, confirmó a la diaria Limber Santos, director del departamento de Educación Rural del CEIP.*
Q2: *¿Cuándo vuelven las clases presenciales a todas las escuelas?* A2: –not found in the text–

Table 2: Example of a short text that could be found in the corpus, and two possible questions for the text. Q1 has the answer 547, found in the text, but Q2 does not have an answer in the text.

one answer. Because of this, when there are multiple answers for a question, the metrics evaluate the answer candidate provided by the system against all the possible answers, and get the maximum value.

5 Competition

The competition was run in two phases: a development phase, for which we released the training dataset with annotations and development dataset without annotations; and an evaluation phase, for which we released the annotations of the development dataset and a test dataset without annotations. Participants could train their models using other available corpora, such as the Spanish version of SQuAD or NewsQA.

5.1 Description of the systems

Eighteen participants registered for the competition in our Codalab site, eight of them submitted results for the development phase (73 submissions in total), and seven of them submitted results for the evaluation phase (46 submissions in total). All of the participants that sent results in the evaluation phase used BERT-like models, analyzing if fine-tuning them with proper data improved their performance.

The language model most commonly used by the participants was RoBERTa for Spanish, trained with the corpus from the Biblioteca Nacional de España[7]. BETO[8], multilingual RoBERTa[9], multilingual BERT[10], and distill BERT for Spanish[11] were also used.

The corpora used, in addition to the QuALES corpus, were SQuAD 2.0, NewsQA and SQAC (Spanish versions).

The participant `smaximo` (Máximo, 2022) followed a curriculum learning strategy consisting of fine-tuning BETO and RoBERTa for Spanish on a series of QA datasets. The author found out that the top performance was achieved using RoBERTa first trained on SQAC, then on the Spanish version of SQuAD (SQuAD-ES-v2) and finally on the QuALES corpus.

[7] https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne

[8] https://github.com/dccuchile/beto

[9] https://huggingface.co/docs/transformers/main/en/model_doc/roberta

[10] https://github.com/google-research/bert/blob/master/multilingual.md

[11] https://huggingface.co/mrm8488/distill-bert-base-spanish-wwm-cased-finetuned-spa-squad2-es

The participant `alvarory` (Rodrigo and Peñas, 2022) tried three main approaches. In the first one they fine-tuned RoBERTa for Spanish (base and large versions) and BETO for 10 epochs on the QuALES training set with datasets containing both training and development splits of the task, for a total of 1,800 question-answer pairs. For the second approach they used even more data than the available in QuALES, in order to study the transferability among different datasets when using two pretrained models: RoBERTa and multilingual BERT. The third approach was based on combining different models for returning a single output using two voting schemes.

The participant `avacaondata` (Vaca-Serrano, 2022) addresses extractive QA through an ensemble system composed of three large pre-trained language models in Spanish: MarIA-base, MarIA-large and RigoBERTa. These models were fine-tuned on data from the Spanish version of SQuAD (SQUAD-ES-v2), a Spanish version of NewsQA, generated by the author, and QuALES. The best model is an ensemble that gives scores to each answer based on multiple criteria such as the number of models that predict it and the models' scores. The final predictions were performed by aggregating the output of the resulting models, referred as a meta-ensemble. A number of ensemble strategies were tried, where finally *Grouped Score Aggregation* perform best. This strategy consists on selecting the answer by the count of each answer multiplied by a scaling factor based on the validation scores of the models.

The participant `Bernardo` fine-tuned RoBERTa for Spanish for 3 epochs using th está hablando.e train subsets of SQAC, SQUAD-ES-v2 and QuALES. `ichramm` performed experiments using RoBERTa for Spanish pretrained with the SQAC corpus, and distill-BERT pretrained with SQUAD-ES-v2. His submitted outputs were calculated using the RoBERTa model, fine-tuned on QuALES. He also experimented including the NewsQA version for Spanish, obtaining lower results (one point less in each metric). The participant `gberger` also used the distill-BERT model.

`sebastianvolti` ranked first in both metrics of the competition. He reached his top performance using XML-RoBERTa, a multilingual model, pretrainend with SQUAD-ES-v2 and fine tuned using the QuALES corpus. He also tested a model that included a fine tuning stage with 2,000 examples from the Spanish translation of the NewsQA corpus, prior to fine tuning with the QuALES corpus, which yielded slightly lower results.

5.2 Results

We show the best result for each user for each metric. Please notice that the best exact match and F1 scores might have been obtained in different submissions by the same user.

Table 3 shows the best exact match scores for each user:

User	EM
sebastianvolti	0.5349
ichramm	0.4677
smaximo	0.4598
Bernardo	0.4427
avacaondata	0.3992
gberger	0.3715
alvarory	0.3175

Table 3: Results for the exact match metric.

Table 4 shows the best F1 overlap scores for each user.

User	F1
sebastianvolti	0.7282
Bernardo	0.6159
smaximo	0.6142
avacaondata	0.5877
ichramm	0.5581
gberger	0.4500
alvarory	0.4293

Table 4: Results for the overlap F1 metric.

As can be seen in the tables, the best results achieved (F1: 0.73 and EM: 0.53) are far from those reported on the SQuAD corpus for English on the official SQuAD site (F1: 0.93 and EM: 0.91). Our task differs from what is reported there in that the evaluation texts belong to a specific domain (news about the Covid-19 pandemic), and also in the size of the context provided to search for the answers. In our case, the context is a complete news article, which are longer than the contexts included in the SQuAD dataset.

The best results were obtained by `sebastianvolti`, whose best model is based on RoBERTa pretrained on SQuAD 2.0, fine tuned on the QuALES corpus, and was the only participant who used the multilingual

version of RoBERTa, four other participants used RoBERTa for Spanish (trained on the BNE corpus).

Also note that none of the systems have reached the inter-annotator agreement levels, both for EM and F1, although for F1 the best submission by `sebastianvolti` is the closest by around 5 %.

6 Conclusions

We presented the results of the QuALES competition on Question Answering Learning from Examples in Spanish. Seven participants submitted systems to the competition, and the best systems achieved 0.53 in exact match and 0.73 in average F1 overlap.

The extractive Q&A task, although showing very good results on the main available benchmark, the SQuAD corpus, still presents great challenges when working with different data and searching for answers in larger contexts.

The QuALES corpus, despite its rather small size, provided significant improvements in training, complementing other larger corpora taken as a base, mainly SQuAD (Spanish version) and SQAC.

References

Baudiš, P. and J. Šedivý. 2015. Modeling of the question answering task in the yodaqa system. In J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. Jones, E. San Juan, L. Capellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 222–228, Cham. Springer International Publishing.

Carrino, C. P., M. R. Costa-jussà, and J. A. Fonollosa. 2019. Automatic spanish translation of the squad dataset for multilingual question answering. *arXiv preprint arXiv:1912.05200*.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:1–10.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gutiérrez Fandiño, A., J. Armengol Estapé, M. Pàmies, J. Llop Palao, J. Silveira Ocampo, C. Pio Carrino, C. Armentano Oller, C. Rodriguez Penagos, A. Gonzalez Agirre, and M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Joshi, M., E. Choi, D. S. Weld, and L. Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension.

Jurafsky, D. and J. H. Martin. 2021. Speech and language processing. 3rd edition draft. *US: Prentice Hall*.

Kwiatkowski, T., J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al. 2019a. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Kwiatkowski, T., J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, M. Kelcey, J. Devlin, K. Lee, K. N. Toutanova, L. Jones, M.-W. Chang, A. Dai, J. Uszkoreit, Q. Le, and S. Petrov. 2019b. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Le, Q. and T. Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Manning, C., P. Raghavan, and H. Schütze. 2010. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103.

Min, S., J. Boyd-Graber, C. Alberti, D. Chen, E. Choi, M. Collins, K. Guu, H. Hajishirzi, K. Lee, J. Palomaki, et al. 2021. Neurips 2020 efficientqa competition: Systems,

analyses and lessons learned. In *NeurIPS 2020 Competition and Demonstration Track*, pages 86–111. PMLR.

Min, S., V. Zhong, R. Socher, and C. Xiong. 2018. Efficient and robust question answering from minimal context over documents. *arXiv preprint arXiv:1805.08092*.

Máximo, S. 2022. Supervised domain adaptation for extractive question answering in spanish.

Nakov, P., D. Hoogeveen, L. Màrquez, A. Moschitti, H. Mubarak, T. Baldwin, and K. Verspoor. 2019a. Semeval-2017 task 3: Community question answering. *arXiv preprint arXiv:1912.00730*.

Nakov, P., L. Màrquez, W. Magdy, A. Moschitti, J. Glass, and B. Randeree. 2019b. Semeval-2015 task 3: Answer selection in community question answering. *arXiv preprint arXiv:1911.11403*.

Peñas, A., P. Forner, Á. Rodrigo, R. Sutcliffe, C. Forăscu, and C. Mota. 2010. Overview of respubliqa 2010: Question answering evaluation over european legislation. In *CLEF*.

Peñas, A., P. Forner, R. Sutcliffe, Á. Rodrigo, C. Forăscu, I. Alegria, D. Giampiccolo, N. Moreau, and P. Osenova. 2009. Overview of respubliqa 2009: Question answering evaluation over european legislation. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 174–196. Springer.

Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

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

Rodrigo, A. and A. Peñas. 2022. Uned@quales 2022: Testing the performance of transformer-based language models for spanish question-answering.

Seo, M., A. Kembhavi, A. Farhadi, and H. Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Trischler, A., T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Vaca-Serrano, A. 2022. Adversarial question answering in spanish with transformer models.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xiong, C., V. Zhong, and R. Socher. 2017. Dcn+: Mixed objective and deep residual coattention for question answering. *arXiv preprint arXiv:1711.00106*.

Yu, L., K. M. Hermann, P. Blunsom, and S. Pulman. 2014. Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632*.