Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 281-287 recibido 05-07-2022 revisado 18-07-2022 aceptado 21-07-2022

Overview of ReCoRES at IberLEF 2022: Reading Comprehension and Reasoning Explanation for Spanish

Overview de ReCoRES en IberLEF 2022: Comprensión de Lectura y Explicación de Razonamiento en Español

Marco Antonio Sobrevilla Cabezudo[1], Diego Diestra[2], Rodrigo López[2], Erasmo Gómez[2], Arturo Oncevay[3], Fernando Alva-Manchego[4]
[1]University of São Paulo
[2]Department of Engineering, Pontificia Universidad Catolica del Perú
[3]School of Informatics, University of Edinburgh
[4]Cardiff University
msobrevillac@usp.br, {ddiestra, a20112387, hector.gomez}@pucp.pe,
a.oncevay@ed.ac.uk, alvamanchegof@cardiff.ac.uk

Abstract: This paper presents the ReCoRES task, organized at IberLEF 2022, within the framework of the 38th edition of the International Conference of the Spanish Society for Natural Language Processing. The main goal of this shared-task is to promote the task of Reading Comprehension and Verbal Reasoning. This task is divided into two sub-tasks: (1) identifying the correct alternative in reading comprehension questions and (2) generating the reasoning used to select an alternative. In general, 3 teams participated in this event, mainly proposing transformer-based neural models in conjunction with additional strategies. The results of this event, insights and some challenges are presented, opening a range of possibilities for future work.
Keywords: Reading Comprehension, Reasoning Explanation, Spanish.

Resumen: Este artículo presenta la tarea ReCoRES, organizada en IberLEF 2022, en el marco de la 38 edición de la Conferencia Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural. El objetivo de esta tarea es promover la tarea de Comprensión de Lectura y Razonamiento Verbal. Esta tarea es dividida en dos sub-tareas: (1) la identificación de la alternativa correcta en preguntas de comprensión de lectura y (2) la generación del razonamiento usado para seleccionar una alternativa. En general, 3 equipos participaron de este evento proponiendo mayormente modelos neuronales basados en transformers con algunas estrategias adicionales. Los resultados de este evento así como aprendizajes y algunos desafíos son presentados, abriendo un abanico de posibilidades como trabajos futuros.
Palabras clave: Comprensión de Lectura, Explicación del Razonamiento, Español.

1 Introduction

Question Answering (QA) consists of returning an accurate and short answer given a Natural Language question. According to Rogers et al. (2020), QA can be approached from two main perspectives: Open QA, in which responses are recovered from several sources such as Web pages and knowledge bases, and Reading Comprehension (RC), where the answer is recovered from a single document.

RC datasets are classified into three categories according to their answer type: (1) span-selection datasets, where the text explicitly includes the answer, (2) multiple-choice datasets, where systems have to select an answer from a list of candidates; and (3) freeform answers dataset, where answers are written in freeform. Most RC datasets are in the first category, with the most popular being SQuAD (Rajpurkar et al., 2016). An explicit limitation of these span-selection datasets is that they can only target information explicitly mentioned in the text and often get solved with shallow lexical match-

ISSN 1135-5948. DOI 10.26342/2022-69-25

ing (Rogers et al., 2020).

Using a multiple-choice dataset is a common and realistic way to measure reading comprehension in humans (Echegoyen, Álvaro Rodrigo, and Peñas, 2020). In addition, Rogers et al. (2020) point out that multiple-choice is a better format to assess language understanding of automatic systems. It is because it requires a high degree of textual inference and the development of strategies for selecting the correct answer.

For English, there are diverse Multiple-Choice QA datasets, such as RACE (Lai et al., 2017), Entrance Exams (Peñas et al., 2011) and QuAIL (Rogers et al., 2020). However, that is not the case for most languages. For Spanish, in particular, there are two QA datasets available: SQuAD-es (Carrino, Costa-jussà, and Fonollosa, 2020) and Entrance Exams (EE) (Peñas et al., 2011). However, these datasets present some limitations originated by the nature of the dataset or some aspects like the size. For example, SQuAD-es is a span-based QA dataset, i.e., the answers are included in the text explicitly. In the case of EE, it is a multiple-choice QA dataset in which, even though questions demand a certain level of reasoning, the dataset size is quite small (43 texts and 191 questions), constraining the exploration of current State-of-the-Art approaches.

In order to contribute to the development of research in Question-Answering/Reading Comprehension for Spanish, this shared-task aims to:

- Introduce a new and more extensive multiple-choice QA dataset for Spanish based on university entrance examinations, where questions aim to evaluate humans instead of computers and include extra information about the reasoning used to choose an alternative.
- Evaluate multiple-choice question answering, and reasoning generation approaches on this dataset.

2 Task Description

This shared-task consists of two sub-tasks:

- Sub-task 1 - Machine Reading Comprehension: given a text, a question, and a set of candidate answers, a system must select the correct answer.
- Sub-task 2 - Reasoning Explanation: given a text and a question, a system must generate an explanation for its answer selection

3 Dataset

The dataset used in this shared-task was extracted from actual university entrance examinations provided by Peruvian institutions that train students for entrance examinations and includes diverse topics and question types that require a certain level of reasoning. Source documents that compose the dataset were initially available in PDF format. This way, we built the dataset by applying two strategies: (1) using an OCR to convert the PDF documents to TXT format and then manually correcting them to fix possible OCR problems, and (2) transcripting the PDF files. Eight collaborators and two organization committee members performed manual revision and transcription.

The whole dataset comprises 439 texts, and 1,822 questions with 2-7 candidate answers each, divided into training, development, and test sets with 257 texts (1,073 questions), 91 texts (363 questions), and 91 texts (386 questions), respectively. Additionally, each question-answer pair instance includes a short explanation as reasoning support for choosing a candidate answer [1].

Figure 1 shows an example of a long text, a question with five alternatives, and the corresponding reasoning. It is worth noting that texts are long, and questions are not described most typically -using question markers and wh-questions; instead, these are described as a sentence that needs to be completed.

4 Experimental Setup

4.1 Baseline

We use two baselines for sub-task 1. The first consists of randomly choosing an answer among the alternatives for each question, and the second is a BERT-based baseline[2], similar to the one used by Rogers et al. (2020). It works this way: for each answer option, the context, question, and choice are joined and

[1]The dataset is available at https://github.com/ddiestra/mrc-dataset.

[2]We use the BERT model available at https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased.

Text: "El trabajo es en primer término un proceso entre la naturaleza y el hombre, proceso en que este realiza, regula y controla mediante su propia acción su intercambio de materias con la naturaleza. En este proceso, el hombre se enfrenta como un poder natural con la materia de la naturaleza. Pone en acción las fuerzas naturales que forman su corporeidad, los brazos y las piernas, la cabeza y la mano, para de ese modo asimilarse, bajo una forma útil para su propia vida, las materias que la naturaleza le brinda. Y a la par que de ese modo actúa sobre la naturaleza exterior a él y la transforma, transforma su propia naturaleza, desarrollando las potencias que dormitan en él y sometiendo el juego de sus fuerzas a su propia disciplina. Aquí no vamos a ocuparnos de las primeras formas de trabajo, formas instintivas y de tipo animal. Aquí, partimos del supuesto del trabajo plasmado ya bajo una forma en la que pertenece exclusivamente al hombre. Una araña ejecuta operaciones que semejan a las manipulaciones del tejedor, y la construcción de los panales de las abejas podría avergonzar, por su perfección, a más de un maestro de obras. Pero, hay algo en que el peor maestro de obras aventaja, desde luego, a la mejor abeja, y es el hecho de que, antes de ejecutar la construcción, la proyecta en su cerebro. Al final del proceso de trabajo, brota un resultado que antes de comenzar el proceso existía ya en la mente del obrero; es decir, un resultado que tenía ya existencia ideal. El obrero no se limita a hacer cambiar de forma la materia que le brinda la naturaleza, sino que, al mismo tiempo, realiza en ella su fin, fin que él sabe que rige como una ley las modalidades de su actuación y al que tiene necesariamente que supeditar su voluntad. Y esta supeditación no constituye un acto aislado. Mientras permanezca trabajando, además de esforzar los órganos que trabajan, el obrero ha de aportar esa voluntad consciente del fin a que llamamos atención, atención que deberá ser tanto más reconcentrada cuanto menos atractivo sea el trabajo, por su carácter o por su ejecución, para quien lo realiza, es decir, cuanto menos disfrute de él el obrero como de un juego de sus fuerzas físicas y espirituales."

Question: Medularmente, el autor intenta dilucidar:

Alternatives:

A. las diferencias entre lo instintivo y lo planificado.
B. la naturaleza del trabajo exclusivamente humano.
C. el carácter pernicioso del trabajo en la actualidad.
D. la supremacía de la naturaleza frente a la humanidad.
E. las etapas que componen el proceso productivo.

Answer: B

Reason: El autor busca caracterizar el trabajo humano frente a lo instintivo, señala así que el trabajo humano está supeditado a un fin.

Figure 1: ReCoRES's Example.

used as input, and the output is its probability, and the most likely option is selected as the answer. Among the settings, we train the model for 1 epoch and use a learning rate of 3e-5 with Adam optimizer.

The baseline for sub-task 2 is a T5-based one that receives the text and the question as inputs and returns the reason[3]. In addition, we evaluate a two-stage approach. Firstly, we select the two most important sentences[4] for a specific question according to cosine similarity. [5]. Then we train a T5-based model, similar to the first baseline. The parameters used were: input length and output length o 512 and 100 tokens, respectively, a learning rate of 0.003 with Adafactor optimizer, and a batch size of 8 with gradient accumulation of 4 steps. Besides, we freeze the embedding layer. Finally, we select the model with the best perplexity in the development set after 7 epochs. During prediction, we use a beam size of 5.

4.2 Evaluation

Sub-task 1 is evaluated in two ways. Firstly, we will evaluate the accuracy, i.e., the number of correct answers in relation to the total number of questions. The second measure is c@1 (Peñas and Rodrigo, 2011), used at CLEF (Rodrigo et al., 2015). c@1 is a conservative metric that penalizes incorrect answers, encouraging systems to not choose an answer unless they are certain.

Sub-task 2 is evaluated in two ways as

[3]We use the T5 model available at https://huggingface.co/flax-community/spanish-t5-small.

[4]We used the two most important sentences because it produced the best results in the development set.

[5]This strategy is inspired by query-based automatic summarization (Hovy, 2005).

well. The first one will consist of running automatic semantic metrics BERTScore (Zhang et al., 2020) to measure the similarity between the generated explanation and its manual reference. We will use this metric instead of classical BLEU (Papineni et al., 2002), or METEOR (Banerjee and Lavie, 2005) because "reasons" can be open and diverse. The second one is a manual evaluation of three quality criteria:

- Accuracy, to measure how accurate is the output system in relation to the original output;
- Fluency (Howcroft et al., 2020), that measures the degree to which a text "flows welländ is not e.g. a sequence of unconnected parts.
- Readability (Howcroft et al., 2020), that measures if the output system is understandable or easy to read.

To perform the manual evaluation, we recruit some crowdworkers. In particular, these were undergraduate students who had experience in this task (Reading Comprehension). The crowdworkers were guided to rate each criteria using an interval of 1-5, being 1 the worst and 5 the best.

5 Participants

In this edition, 3 teams registered on the task and submitted results. However, two of them presented working notes describing their systems. The following is a brief summary of the final proposals submitted:

5.1 MRCPUCP

This team only participated in the sub-task 1[6]. They proposed a BERT-based approach in which all text, alternatives, and reasoning are concatenated and used as input, and the output is one of the alternatives. They used BETO as BERT-based model for Spanish (similar to the baseline) and finetune the model on the dataset they built.

5.2 SADA (Baggetto et al., 2022)

This team only participated in the sub-task 1. The authors explore using encoder models, generative models, clue generation systems, and dataset expansion. In experiments, the best model was a pre-trained multilingual T5 model finetuned on an expanded multilingual dataset.

[6]This team did not present working notes for the present shared-task.

Sub-task 1		
Team	**Accuracy**	**c@1**
Baseline (Random)	0.2514	0.2514
Baseline (BERT)	0.1917	0.1917
Baseline (BERT) + Threshold	0.0492	0.0896
Versae & Nandezgarcia	0.4067	0.4067
SADA	0.7254	0.7254
MRCPUCP	**0.7591**	**0.7591**
Sub-task 2		
Team	**BERTScore**	
Baseline T5	0.6579	
T5 - 2 sentences	0.6652	
Versae & Nandezgarcia	**0.6867**	

Table 1: Results of Automatic Evaluation.

5.3 Versae & Nandezgarcia (De la Rosa and Fernández, 2022)

This team participated in both sub-tasks. The authors tested several methods for classic fine-tuning of encoder-only language models for the task of reading comprehension and a zero-shot approach for reasoning explanation using a decoder-only model.

6 Results and Discussion

Table 1 presents the results for the sub-task 1 and sub-task 2. For sub-task 1 (Reading Comprehension), the best performance was obtained by the MRCPUCP team, and the SADA team obtained the second-best one, only 3 points lower than the first one.

It is worth noting that all teams used pretrained models such as BERT (Devlin et al., 2019) or T5 (Raffel et al., 2020) in conjunction with some additional strategies. Among the strategies, we can highlight the use of reasoning as part of the input and its helpfulness in getting the correct answers in the reading comprehension task. On the other hand, multilingual information has proven to be helpful, even when the domains are different.

Concerning sub-task 2, the best performance in the automatic evaluation was obtained by the work of Versae & Nandezgarcia, being almost 2 points higher than the strong T5-based baseline. It is worth noting that the winning proposal used a zero-shot approach, i.e., no training data of this task was used for learning to generate the reasoning.

Due to input texts in our dataset being long, we wonder how much do text length influence the performance? To verify it, we

divide the test set in text subsets according to its length, as shown in the X-axis in Figures 2 and 3.

Figure 2 shows how the accuracy changes according to the text length for all proposals (baseline is the BERT-based one). We can note that, as was expected, the performance decreases when the texts are longer, except for the cases where the length is higher than 500 tokens. This result is suspicious as most proposals were BERT-based models. Thus, the maximum length was defined as 512 tokens. However, the proposal of SADA uses a T5-based model that can deal with these lengths. In the case of the longest texts (between 850 and 900 tokens), we must note that the performance was almost 0.25 because the models usually chose an alternative by chance, and it was correct for all questions that had the same alternative as correct.

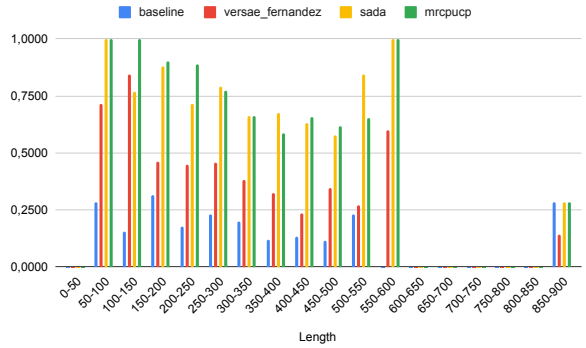

Figure 2: Analysis of the performance on the Machine Comprehension task according different text lengths.

Figure 3 shows how BERTScore changes according to the text length for all proposals. We note that even when the values obtained by Versae & Nandezgarcia are a bit higher in all subsets, these are almost the same (a bit higher than 0.60). These results can suggest that models can deal with different text lengths in the same way or that metric is not good enough to determine what is the best proposal. However, a deeper study is necessary to determine the actual reason for getting these results.

Finally, Table 2 presents the human evaluation results. It shows that fluency and readability achieve similar scores (almost 4) for all proposals, being a bit better for the proposal of Versae & Nandezgarcia. This is expected as all models are based on big language models that can usually generate fluent and readable texts. In the case of accuracy, we can see that the proposal of Versae & Nandezgarcia obtained the best results. However, results are still lower than 3, proving that this task is harder and the automatic evaluation metric could not be suitable.

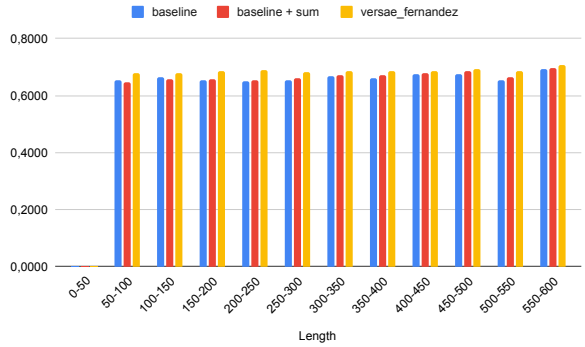

Figure 3: Analysis of the performance on the Reasoning Explanation task according different text lengths.

	Accuracy	Fluency	Readability
Baseline T5	1.08 ±0.40	4.04 ±1.16	3.95 ±1.26
T5 - 2 sentences	1.20 ±0.51	4.08 ±1.07	3.93 ±1.22
Versae & Nandezgarcia	2.35 ±1.39	4.33 ±0.90	4.33 ±0.92

Table 2: Human Evaluation. Accuracy, Fluency and Readability were rated in an interval of 1-5. Results are shown in terms of mean ± standard deviation.

7 *Conclusion*

We presented the first edition of the ReCoRES task at IberLEF, including two sub-tasks: reading comprehension and reasoning explanation.

In general, three teams participated in this shared-task: three for sub-task 1 and one for sub-task 2. However, only two teams sent their working notes. All proposals were based on pre-trained language models with some additional strategies.

Overall, the winner of sub-task 1 was the MRCPUCP team, and the winner of sub-task 2 was the Versae-Nandezgarcia team. About the results, some interesting findings about the helpfulness of incorporating reasoning information and multilingual datasets in the reading comprehension task and the need to use more suitable metrics and other strategies to deal with the reasoning explanation task as this one has proven to be complicated.

As future work, we plan to extend the cur-

rent corpus for both sub-tasks and annotate different question types according to the taxonomy proposed by Rogers et al. (2020) to verify what are the actual abilities of pretrained language models. Besides, we plan to annotate text segments that explain the reasoning to build an extractive reasoning explanation dataset instead of an abstractive one like the one used in this shared-task.

Acknowledgements

The authors acknowledge the support of the undergraduate students at the department of Software Engineering at the Universidad Nacional Mayor de San Marcos in the manual evaluation.

References

Baggetto, P., S. Ramos, J. García, and J. R. Navarro. 2022. Study on text comprehension and MCQA in spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*, A Coruña, Spain. CEUR Workshop Proceedings.

Banerjee, S. and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Carrino, C. P., M. R. Costa-jussà, and J. A. R. Fonollosa. 2020. Automatic Spanish translation of SQuAD dataset for multi-lingual question answering. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5515–5523, Marseille, France, May. European Language Resources Association.

De la Rosa, J. and A. Fernández. 2022. Zero-shot Reading Comprehension and Reasoning for Spanish with BERTIN GPT-J-6B. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*, A Coruña, Spain. CEUR Workshop Proceedings.

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 American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Echegoyen, G., Álvaro Rodrigo, and A. Peñas. 2020. Cross-lingual training for multiple-choice question answering. *Procesamiento del Lenguaje Natural*, 65(0):37–44.

Hovy, E. 2005. Text summarisation. *The Oxford Handbook of computational linguistics*, pages 583–598.

Howcroft, D. M., A. Belz, M.-A. Clinciu, D. Gkatzia, S. A. Hasan, S. Mahamood, S. Mille, E. van Miltenburg, S. Santhanam, and V. Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland, December. Association for Computational Linguistics.

Lai, G., Q. Xie, H. Liu, Y. Yang, and E. Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September. Association for Computational Linguistics.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Peñas, A., E. H. Hovy, P. Forner, Á. Rodrigo, R. F. E. Sutcliffe, C. Forascu, and C. Sporleder. 2011. Overview of QA4MRE at CLEF 2011: Question answering for machine reading evaluation. In V. Petras, P. Forner, and P. D. Clough, editors, *CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands*, volume 1177 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Peñas, A. and A. Rodrigo. 2011. A simple measure to assess non-response. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1415–1424, Portland, Oregon, USA, June. Association for Computational Linguistics.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.

Rodrigo, Á., A. Peñas, Y. Miyao, E. H. Hovy, and N. Kando. 2015. Overview of CLEF QA entrance exams task 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.

Rogers, A., O. Kovaleva, M. Downey, and A. Rumshisky. 2020. Getting closer to ai complete question answering: A set of prerequisite real tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8722–8731, Apr.

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.