URJC-Team at PROFE2025@ IberLEF: Deep Language Comprehension Using Transformers Voting Architecture

Miguel Ángel Rodríguez-García[1], Raúl Cabido[2], Michel Maes-Bermejo[2] and Soto Montalvo[2]

[1]NLP IR Group, Dpto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), Juan del Rosal 16, Madrid 28040, Spain

[2]Dpto. Informática y Estadística, Universidad Rey Juan Carlos (URJC), Calle Tulipán s/n, Móstoles 28933, Madrid, Spain

Abstract

Machine Reading Comprehension (MRC) is a challenging task focused on developing systems that can read text and understand its meaning, which is crucial for question-answering systems. This area has gained significant attention from the research community, leading to the emergence of shared evaluation campaigns that promote exciting challenges in the field. In this context, the IberLEF 2025 includes the PROFE task, which centers on deep language comprehension. In this work, we describe the system proposed for one of the tasks in this challenge: matching answers. Our approach is primarily based on fine-tuning three Spanish language models RoBERTA, BETO, and BERT and applying a voting technique to determine the best match for each question. For training the models, we used external training data from a public dataset for Spanish question answering.

Keywords

Transformers, Deep comprehension, Question answering, Answer extraction

CEUR-WS.org/Vol-4098/PROFE_paper5.pdf

1. Introduction

One of the main challenges of natural language processing is that machines can be able to understand text as well as humans. Machine Reading Comprehension (MRC) is an essential task in evaluating natural language understanding. One common method for evaluating someone's understanding of text is to give a passage, then they must answer the questions correctly. Then, for MRC is necessary to training a machine to predict the answer to a question given a relevant context [1]. If the machine can answer the question correctly, we can conclude that it is capable of understanding the context and inferring the question from it [2].

Various works have focused on MRC, providing new datasets to encourage research [3], as well as benchmarks designed to thoroughly assess the MRC capabilities of LLMs [4], among others. In this line, the PROFE task [5] at IberLEF 2025 [6] focuses on deep language comprehension including three subtasks, where each subtask contains several exercises of the same type. On the first subtask (multiple choice) given a multiple-choice question, systems must select the correct answer among the candidates. On the second subtask (matching), each exercise contains two sets of texts and systems must find the text in the second set that best matches the first set. Finally, on the third subtask (filling-in-the-gap) each exercise contains a text with several gaps corresponding to textual fragments and systems must determine the correct position for each fragment.

This work presents a system developed for the second subtask. We approach the challenge combining different models designed for Spanish Question Answering tasks.

The remainder of the article is organized as follows: Section 2 presents a focused study of the approaches analyzed for building the proposed system. Section 3 offers an overview of the dataset and methods used in the campaign. Section 4 discusses the results obtained. Finally, Section 5 outlines the conclusions drawn from our participation in the shared evaluation campaign, and suggests how improve the proposal for future research.

IberLEF 2025, September 2025, Zaragoza, Spain

miguelangel.rodriguez@lsi.uned.es (M. Rodríguez-García); raul.cabido@urjc.es (R. Cabido); michel.maes@urjc.es (M. Maes-Bermejo); soto.montalvo@urjc.es (S. Montalvo)

0000-0001-6244-6532 (M. Rodríguez-García); 0000-0002-9178-9224 (R. Cabido); 0000-0002-8138-9702 (M. Maes-Bermejo); 0000-0001-8158-7939 (S. Montalvo)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-10

2. Related Work

In this section, we present a brief analysis of the Machine Reading Comprehension task within the teaching domain. We outline various approaches that have been foundational in developing the architecture for our participation in the PROFE task.

Zoukagh [7] focuses on designing and implementing BERThiz, a model aimed at understanding contextual cues and generating accurate responses for masked inputs. The proposed model surpasses other models within the same category and language. For Multi-Language Question Answering (MLQA) task, the proposed model obtain the same results as BETO model. In [8] the authors present an study explored the challenging task of automatically generating Spanish multiple-choice question using multilingual language models.

Ruiz et al. [9] proposes a question-answering system designed to assist workers in the manufacturing industry in requesting information from technical manuals. The authors manually annotated these manuals and conducted experiments with various Spanish language models (five in Spanish and 1 Multilingual). The results show that the proposed system will enhance the efficiency of the operators, as it allows them to perform routine tasks without physically consulting the manual.

More general approach is presented in [10]. This work is focused on analyzing the performance of Spanish language models across various natural language tasks. This study examined the convergence of Spanish sequence-to-sequence models while considering different downstream tasks, such as summarization, splitting and rephrasing, dialogue, and question answering, among others.

We have explored additional approaches to the state-of-the-art not presented in this section, and their analysis was crucial in pushing us to employ language models for the PROFE task.

3. Materials and methods

In this section, we discuss the models used to structure the system's architecture and the resources utilized during the training stage.

3.1. Dataset

Training is a crucial stage when adapting a deep learning model built for one task to work on another. Since the number of examples provided in the PROFE task was limited, we decided to seek out similar datasets related to question-answering tasks. From these datasets, we adapted the information to fit the format required for our specific task in order to train the selected models.

For this purpose, we chose the Spanish Question-Answering Corpus (SQAC) dataset [1], which contains 6,247 contexts and 18,817 questions, each with corresponding answers rated on a scale from 1 to 5. The dataset is divided into three files: training, development, and testing, each varying in size. We specifically selected the training file detailed in this proposal. To create the synthetic dataset for training, we defined the following procedure. This file is structured as a list of paragraphs, each consisting of several attributes: i) "context," which represents the context of the query; ii) "qas," which includes a list of three attributes: question, id, and answer. The "question" refers to the query itself, "id" represents its unique identifier, and "answer" provides both the position of the answer and the corresponding sentence.

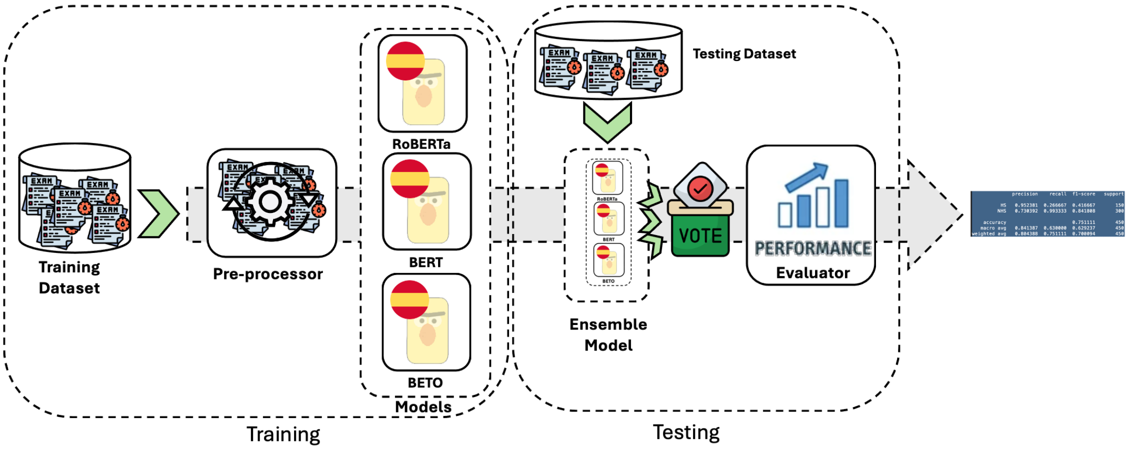
We mapped this information into a similar structure provided by the organizers. For each question and its answer, we selected four different contexts, labeled A, B, C, and D, from the dataset. We ensured that one of these contexts contained the correct answer. This process was repeated for each question available in the dataset. Consequently, we compiled hundreds of questions into the synthetic dataset.

[1] https://huggingface.co/datasets/PlanTL-GOB-ES/SQAC

3.2. Architecture of the system

The primary approach taken to tackle the task was to utilize pre-trained models for question-answering tasks in the Spanish language. Initially, we tested several language models, including DistilBERT, BERT, RoBERTa in its various versions, and T5, among others. We then ranked these models based on their performance using the prepared dataset during training. From this evaluation, we selected the top three models to form the proposed system.

The models that achieved the best results in our initial experiments were BETO [2], BERT MMG/bert-base-spanish-wwm-cased-finetuned-sqac-finetuned-squad, and RoBERTa [3]. We combined these models using a voting strategy: for each question posed, each model voted for the best matching answer. Ultimately, the answer that received the most votes was selected as the final response to link with the question. Figure 1 illustrates the architecture of the designed system.



```
              precision    recall  f1-score   support

          HS   0.952381  0.266667  0.416667       150
         NHS   0.730392  0.993333  0.841808       300

    accuracy                       0.751111       450
   macro avg   0.841387  0.630000  0.629237       450
weighted avg   0.804388  0.751111  0.700094       450
```

Figure 1: Architecture of the system designed

The proposed architecture, as shown in Figure 1, consists of a pipeline with three main stages. The first stage is the preprocessing phase. The selected dataset for training did not match the format provided by the organizers, so it needed to be preprocessed. This involved reorganizing its structure to group the contexts, questions, and answers into a user-friendly data format that facilitates the training stage; the second stage is the training phase. The three selected models were trained using the dataset compiled from the previous stage; and, finally, the third phase focuses on providing the final prediction by managing a voting system of the three models integrated into the ensemble.

4. Results

This section outlines the experiments conducted for the shared evaluation campaign. The first step involved selecting language models trained specifically on the Spanish Question-Answering Corpus. We configured each model using the following settings: 3 epochs, a learning rate of 5e-5, a weight decay of 0.01, and a batch size of 16. Subsequently, we used our dataset to independently train and evaluate each model. The dataset was split into three parts: 70% for training and 20% for testing and 10% for validation. Table 1 presents the performance of each model over de testing in terms of accuracy.

Our results demonstrate a high level of performance for each model, approaching values very close to 1. This leads us to believe that creating an ensemble model could yield significantly better results. This formed the basis of our hypothesis for building the ensemble model.

Table 2 presents the official results achieved by the ensemble model on the test set provided by the organizers.

[2] Josue/BETO-espanhol-Squad2

[3] PlanTL-GOB-ES/roberta-base-bne-sqac

Table 1
Preliminary results

Models	Accuracy
BETO	0.99
RoBERTa	0.97
BERT	0,96
Ensemble	0.98

Table 2
Participant results on the test dataset on matching subtask.

Approaches	Task2 Matching(Accuracy)
ezotova 2	95.91
ezotova 1	89.39
ezotova 4	84.31
ezotova 5	71.93
ezotova 3	57.35
djanr2 ID 285662	43.77
djanr2 ID 285677	43.28
URJC_TEAM	**34.6**
djanr2 ID 285670	22.25
djanr2 ID 285674	21.52
djanr2 ID 285672	19.19
ymlopez submissiondeepseek-qwen.zip	4.89
ymlopez submissionqwen-think.zip	4.89
ymlopez submission.zip	4.89
ymlopez submissionqwen3.0.zip	4.89

When comparing the performance of the ensemble model using our dataset and the ensemble model evaluated with the test provided by the organizers, a significant difference in accuracy becomes apparent. The ensemble model for the given test dropped to 0.40, showcasing a loss of more than half of its accuracy. This stark decline suggests that neither the isolated models nor the ensemble model can generalize the knowledge learned during training to apply it to unknown instances. This scenario indicates a substantial gap between the two datasets: the one we prepared and the one used in the test.

5. Conclusions and future works

The PROFE task challenges researchers to build a system capable of understanding information described in natural language in Spanish and answering three different types of questions: multiple choice, matching, and filling-in-the-gap. In this work, we focused specifically on the matching subtask, which involves linking each context to the corresponding answer based on a given set of contexts and answers.

We fine-tuned three Spanish language models (BETO, BERT, and RoBERTa), and applied a voting ensemble approach to combine their individual predictions. Our low rankings on the leaderboard indicate that there is significant room for improvement.

We are considering several new directions for future work. On one hand, we plan to use multilingual datasets with multilingual models to evaluate their performance. On the other hand, we will dismiss the dataset compilation and exploring language models with few-shot strategies to analyze their convergence. Finally, we want to explore generative large language models from different point of view.

Acknowledgments

Action funded by Universidad Rey Juan Carlos (URJC) under the 2024/2025 Educational Innovation Projects Call, project code PIE24_092.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Z. Chen, R. Yang, B. Cao, Z. Zhao, D. Cai, X. He, Smarnet: Teaching machines to read and comprehend like human, arXiv preprint arXiv:1710.02772 (2017).

[2] Y. Cui, T. Liu, W. Che, Z. Chen, S. Wang, Teaching machines to read, answer and explain, IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2022) 1483–1492.

[3] R. Matthew, C. J. Burges, E. Renshaw, Mctest: A challenge dataset for the open-domain machine comprehension of text, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 193–203.

[4] S. Ma, H. Peng, L. Hou, J. Li, Mrceval: A comprehensive, challenging and accessible machine reading comprehension benchmark, 2025. URL: https://arxiv.org/abs/2503.07144. arXiv:2503.07144.

[5] Á. Rodrigo, S. Moreno-Álvarez, A. Pérez García-Plaza, A. Peñas, R. Agerri, J. Fruns-Jiménez, I. Soria-Pastor, Overview of PROFE at IberLEF 2025: Language Proficiency Evaluation, Procesamiento del Lenguaje Natural 75 (2025).

[6] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.

[7] I. Zoukagh, Bethiz: Precise and complete bert model to fill out questions with correct answers, understanding the context for the spanish language, Authorea Preprints (2024).

[8] D. de Fitero-Dominguez, A. Garcia-Cabot, E. Garcia-Lopez, Automated multiple-choice question generation in spanish using neural language models, Neural Computing and Applications 36 (2024) 18223–18235. URL: https://doi.org/10.1007/s00521-024-10076-7.

[9] E. Ruiz, M. I. Torres, A. del Pozo, Question answering models for human–machine interaction in the manufacturing industry, Computers in Industry 151 (2023) 103988.

[10] V. Araujo, M. M. Trusca, R. Tufiño, M.-F. Moens, Sequence-to-sequence Spanish pre-trained language models, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 14729–14743. URL: https://aclanthology.org/2024.lrec-main.1283/.