BioMedIA: A Complete Voice-to-Voice Generative Question Answering System for the Biomedical Domain in Spanish

Alejandro Vaca Serrano[1], David Betancur Sánchez[1], Alba Segurado[1]
Guillem García Subíes[1], Álvaro Barbero Jiménez[1,2]
[1]Instituto de Ingeniería del Conocimiento, Madrid, Spain
[2]Universidad Autónoma de Madrid, Madrid, Spain
{name.firstsurname}@iic.uam.es

1 Introduction

The objective of this work is to develop a reliable and complete Generative Question Answering (QA) System in Spanish, for the biomedical domain. The need for such kind of system for general users to clarify complex biomedical questions is noticeable, given the existing misinformation and the lack of reliable tools that join multiple sources to form a complete answer about health-related topics. Given the importance of these for society as a whole, and the lack of relevant resources in Spanish, it was considered of general interest to develop a system that could bring together the knowledge located in different sources and make it available to the Spanish-speaking community. Moreover, putting a focus on accessibility, the system should also be fully operated through voice.

2 Background

Up to recently, QA systems were usually built with two pieces (see (Karpukhin et al., 2020)): a) an information retrieval system, based on BM25, TF-IDF, or Sentence Transformers, and b) an extractive QA model, which selects parts of the texts obtained by the piece above and returns them as an answer.

Currently, the existing NLP technologies and resources for English allows creating more advanced solutions, such as Wikipedia Assistant (Blagojevic, 2022), which rely on Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) and Long Form Question Answering (LFQA) (Blagojevic, 2022) models . This, was not previously possible for Spanish, due to the relatively small number of publicly available resources for this language, and in particular for the task of training passage retrieval and generative QA models, in spite of being one of the most spoken languages in the world.

The main contribution of this work is BioMedIA, a LFQA system for the biomedical domain in the Spanish language. This is, to the best of our knowledge, the first time Dense Passage Retrieval (DPR) models have been trained in Spanish with large datasets, and the first time a generative QA model in Spanish has been released. All the codebase is published as open-source, and we also contribute to the NLP community with automated translations to Spanish of the text similarity, QA and LFQA datasets used for training BioMedIA.

3 Methodology

3.1 System architecture

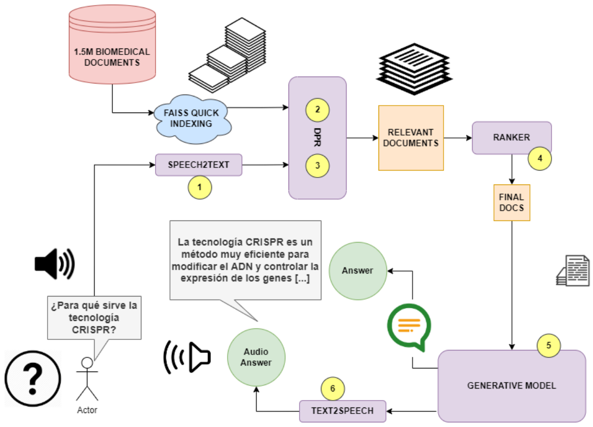

Figure 1: Architecture of BioMedIA.

Figure 1 presents the architecture of the proposed BioMedIA system. Users can input questions through free-form text, or as a voice message that is transcripted to text. The DPR module then encodes the question as an embedding, which is compared against a database of crawled biomedical texts (CoWeSe) (Carrino et al., 2021) with precomputed DPR embeddings. An optimized FAISS index (Johnson et al., 2019) is used for quick retrieval of the most relevant passages. A more fine-grained selection of passages is then performed by a ranker model, which are forwarded to a generative QA model producing the answer in text form. Finally, an audio answer is also generated using a text to speech (T2S) model.

3.2 Datasets

We now describe the datasets used for training the different models of the proposed system. As some of them were available only for the English language, as part of this work we applied the automated translation model marianMT (Tiedemann and Thottingal, 2020) due to its precision-efficiency balance (Junczys-Dowmunt et al., 2018).

3.2.1 DPR datasets

BioAsq_es (translated): translation of a QA corpus for the biomedical domain (Nentidis et al., 2021), created by a team of biomedical experts. As the translation process might alter the wording of answers and related contexts, we developed an alignment algorithm based on sentence tokenization and intersection of the words present in the answer and in the portion of the context that we are evaluating, so that only the paragraph from the context that matches the answer is extracted.

SQAC (Gutiérrez-Fandiño et al., 2022): a QA dataset containing 6,247 contexts and 18,817 questions with their answers, 1 to 5 for each fragment.

SQuAD-ES (Carrino et al., 2019): an automatic translation of the Stanford Question Answering Dataset (SQuAD) v2 (Rajpurkar et al., 2016) into Spanish.

3.2.2 Ranker dataset

MSMarco_es (translated): a Spanish version of msmarco v1 (Nguyen et al., 2016) , a dataset used for text similarity tasks. Further processing was required to sample the queries, as there were some of them with a different ratio of positive and negative labels than the recommended (4 neg and 1 pos) (Reimers and Gurevych, 2019).

3.2.3 LFQA datasets

LFQA_es (translated): a Spanish version of lfqa (Blagojevic, 2022), used for LFQA training.

3.3 Models

3.3.1 Speech to Text model

Arguably, the model holding current State-of-the-Art (SOTA) for English is Wav2Vec2 (Baevski et al., 2020), and although its multilingual version XLSR-53 (Conneau et al., 2020) also works for Spanish, it is not specific for this language. It was also identified that no model trained with big corpora like Multilingual Librispeech (Pratap et al., 2020) was openly available for Spanish. Thus, for this work the large version of XLSR-53 was fine-tuned on Multilingual Librispeech, following the procedure in (Conneau et al., 2020), to conform the speech to text module.

3.3.2 DPR: Dense Passage Retriever

Dense Passage Retriever (DPR) (Karpukhin et al., 2020) is the SOTA passage retrieval model, originally developed in English, consisting of two BERT (Devlin et al., 2018) models, one for encoding passages and the other for encoding questions. For training such a model, authors in the original paper used several extractive QA datasets. For each question, they took the relevant passage (the one containing the answer) as the positive example. For the negative examples, they took 4 in total per each positive one; 3 of them are selected by picking passages relevant to other questions, and one is selected by getting the passage BM25 (Robertson and Zaragoza, 2009) would choose as the most relevant, excluding the positive one. In this work, a Spanish version of DPR is implemented by using the train split of the datasets introduced in 3.2.1, following the hyperparameter settings in (Karpukhin et al., 2020) and BETO (Cañete et al., 2020) as the base model.

3.3.3 Passages Ranker

After relevant passages are selected, BioMedIA ranks them based on relevance to the query, using only the top 5 articles for generating the answer. Three different configurations were used.

Multilingual Sentence Transformer: this was the first option, since no models were available in Spanish for this task. A Sentence Transformer from Sentence-Transformers library (Reimers and Gurevych, 2019) was used.

Monolingual Spanish Cross-Encoder: with the use of Sentence-Transformers library (Reimers and Gurevych, 2019), a Cross-Encoder was trained on MSMarco_es, introduced above, using Roberta-base (Gutiérrez-Fandiño et al., 2022) from the MarIA project as the base model.

Combination of both: there was a great rank distribution disparity between both systems. With the aim to offset each model's bias, their similarity scores are multiplied, thus producing a more reliable rank.

3.3.4 Generative Question Answering Model

For the generative QA part of the system, the LFQA_ES dataset is used. The model input is the

Model	WER
xlsr-53	11.5
ours	**7.3***

Table 1: Word Error Rate (WER) (Ali and Renals, 2018) for Speech to Text models on Multilingual Librispeech test split. Lower is better.

Metric	dpr-squad	dpr-allqa
F1-Macro	0.880	**0.945***
avgrank	0.274	**0.117***

Table 2: Test results on SQUAD-ES for both DPR models. We measure relevant vs not relevant f1 performance (higher is better), and average rank in the ranking task (lower is better).

query plus the most relevant passages to answer it, while the output is the answer. For this type of task, an Encoder-Decoder model architecture is needed; as there is no such monolingual model in Spanish, two different multilingual Encoder-Decoder models were used as base models:

MT5-base-lfqa: the MT5-base (Xue et al., 2020) model was used as the base model. It is a multilingual Encoder-Decoder model trained by Google on the mC4 corpus (Xue et al., 2020).

MBART-large-lfqa: the base model used was MBART-large (Liu et al., 2020), developed by Meta, primarily focused on machine translation, but also suitable for other text generation tasks like the one at hand.

As for the hyperparameters, a similar setting as (Blagojevic, 2022) was used for both models, which are of similar size.

3.3.5 Text to Speech (T2S)

To translate the system output text into speech, Meta's T2S model (Wang et al., 2021) in Spanish was used.

4 Experiments and Results

The standard metrics for each task are used for evaluation.

Speech to Text: as shown in Table 1, our model shows a significant improvement when compared in terms of WER against the XLSR-53 model on the Multilingual Librispeech dataset (Pratap et al., 2020).

Dense Passage Retrieval: as no DPR models were available for Spanish, we trained a strong baseline, denoted as dpr-squad on Table 2, using only the train split of SQUAD-ES, so as to gauge the improvements provided by the extra datasets we prepared, denoted by dpr-allqa on the same Table. Both models, dpr-squad and dpr-allqa, were evaluated (Table 2) using two metrics on the validation set of SQUAD-ES, as this was used as the test set, while a random portion of the train set was used for the development set.

Model	MRR@10
Multiling-SentenceTrans.	0.5891
Roberta-Ranker (ours)	0.6880
Combination of both	**0.6935***

Table 3: Eval results on MSMarco_ES for both Ranker models. Higher is better.

Passages Ranker: Table 3 shows the performance of the Multilingual SentenceTransformer and the Roberta-based ranker introduced in this work in terms of MRR@10 (Mean Reciprocal Rank @ 10) (Craswell, 2009). It can be appreciated that the monolingual model clearly outperforms its multilingual counterpart, in spite of being formed by one encoder instead of two.

Generative Question Answering Model: metrics for both LFQA models on the development set of LFQA dataset can be found at Table 4.

5 Conclusions

In this work a complete LFQA system for the biomedical domain in Spanish was presented. To this end, novel techniques relevant for several information retrieval tasks in Spanish were developed, such as a DPR, a performing Wav2Vec2 model, a ranker model trained on monolingual data and generative QA models. We hope these contributions will aid the Spanish NLP community in reducing the gap to the English language in terms of NLP resources.

Acknowledgements: this work was developed as part of the SomosNLP Spanish Hackathon.

Metric	MT5-base-lfqa	MBART-large-lfqa
Rouge1	**10.291***	0.511
Rouge2	**1.725***	0.004
RougeL	**8.919***	0.511
RougeLSum	**7.987***	0.511

Table 4: Dev results on LFQA_ES for both LFQA models in rouge metrics (Lin, 2004). Higher is better.

References

Ahmed Ali and Steve Renals. 2018. Word error rate estimation for speech recognition: e-WER. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 20–24, Melbourne, Australia. Association for Computational Linguistics.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

Vladimir Blagojevic. 2022. Long-form qa beyond eli5: an updated dataset and approach.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Ona de Gibert Bonet, Asier Gutiérrez-Fandiño, Aitor Gonzalez-Agirre, Martin Krallinger, and Marta Villegas. 2021. Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models.

Casimiro Pio Carrino, Marta R. Costa-jussa, and Jose A. R. Fonollosa. 2019. Automatic Spanish Translation of the SQuAD Dataset for Multilingual Question Answering. *arXiv e-prints*, page arXiv:1912.05200v1.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, abs/2006.13979.

Nick Craswell. 2009. *Mean Reciprocal Rank*, pages 1703–1703. Springer US, Boston, MA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68(0):39–60.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *CoRR*, abs/2004.04906.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *CoRR*, abs/2001.08210.

Anastasios Nentidis, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, and Georgios Paliouras. 2021. Overview of bioasq tasks 9a, 9b and synergy in clef2021. In *Proceedings of the 9th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *Interspeech 2020*. ISCA.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, page arXiv:1606.05250.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Changhan Wang, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Ann Lee, Peng-Jen Chen, Jiatao Gu, and Juan Pino. 2021. fairseq s^2: A scalable and integrable speech synthesis toolkit. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 143–152, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934.