International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 14 (2022) pp. 119-127

Received: 12 January, 2022; Accepted: 3 February, 2022; Publish: April 21, 2022

A BERT-based Question Answering Architecture for Spanish Language

Robert C. Gutierrez Ramos[1], Hugo D. Calderón-Vilca[2] and Flor C. Cárdenas-Mariño[3]

[1] Software Engineering Department, Universidad Nacional Mayor de San Marcos,
Germán Amézaga Street s/n, Lima, Perú
robert.gutierrez1@unmsm.edu.pe

[2] Software Engineering Department, Universidad Nacional Mayor de San Marcos,
Germán Amézaga Street s/n, Lima, Perú
hcalderonv@unmsm.edu.pe

[3] Operations Research Department, Universidad Nacional Mayor de San Marcos,
Germán Amézaga Street s/n, Lima, Perú
fcardenasm@unmsm.edu.pe

***Abstract*: QA systems have had various approaches to achieve their goal of solving naturally formed questions, recent works use state of the art techniques such as neural networks, QA systems in different languages are increasing, as evidenced, they are advancing at different rates, despite the fact that there are efforts to increase research in this type of systems. In this research we analyze the main aspects of the contributions to Question Answering and present an architecture that is capable of answering questions in Spanish. The initial question is received by the system, which may or may not have a document corpus from which to extract the answer, if it does not have a specified document corpus, the Answer Generation module returns the answer to the initial question. The purpose of the system is to provide answers to factoid questions posed by users through a web and mobile platform. BI-LSTM was used for document retrieval and BERT was used to generate the answers. We tested the architecture with ten thousand questions reaching an accuracy of 0.7856. The result improved by entering QA to a more specialized BERT model adapted for the Spanish language, the multilingual version of BERT and the Spanish version of BETO were used.**

***Keywords*:** Artificial intelligence,Information retrieval,Learning systems,Neural networks,Machine learning.

Introduction

QA systems have had various approaches to achieve their goal of solving naturally formed questions. Recent works use state of the art techniques such as neural networks, such is the case [1] who presents a model for QA based on reinforced learning and multi-document summarization techniques. In comparison to other methods such as the use of graphs and RDF like the work of [2] which in contrast focuses more on data graph oriented semantic categorization, so that it breaks down the knowledge and easily deals with the problem of question ambiguity. More recently there is the work of [3], who proposes a QA model for the restricted domain of home service, whereby he proposes a sequence selection tool in conjunction with neural networks and uses information based on information sources such as WikiHow, WikiSmall and Wiki Answer. In the same way, [4] proposes a model that is based on Multi-Layer Transformer Aggregation (MTA), thus reinforcing the representation of the information of his model, finally validating it based on the SQuAD dataset, also of the most famous SQuAD (Stanford Question Answering Dataset) used by [5], [6]. On the other hand, in the proposal of [7] they compare two algorithms for routing questions and answers, applied to group of students software engineering.

Thus, every day there is a growing demand for access to different types and sources of information. Huge repositories of unstructured information are growing all the time, this available information can come from various types of web pages, books, encyclopedias, experts, communities and people asking questions about topics of interest on a daily basis. Usually, the answers to these questions lie within some text of all digital content. This “hidden” knowledge is immersed in a sea of information, the precise answers to which cannot be accessed in all cases [8].

Question and answer (QA) websites addressed this challenge effectively. The human variant of QAs was used effectively over the years. These QA platforms emerged from the early 1990s such as Yahoo Answers or Reddit in the 2000s until today. The main limitation of these QA systems is that they rely on human interaction to help provide the answers. The goal of open domain QA is to automate this process so that computers can do the same as professional information analysts through the use of artificial intelligence [1]. These capabilities are widely used for different purposes, but having as a constant among them all, the acquisition of knowledge, which applied in the sociocultural context of Peru, could be an ally in the general education of the population.

According to UNICEF data [9], more than 90% of primary school-age children are enrolled in school worldwide, and this situation has been maintained for the last 10 years, of

which only 69% successfully complete the school year. It is hoped that the proposed model will help more students complete the school year. A question-and-answer system serves as a first-hand tool for students, who, by resolving their doubts, will better consolidate the knowledge they have acquired. These educational aids could be of great benefit to children's education at the macro level.

Although the advances in QA systems in different languages are increasing, as evidenced, they are advancing at different rates, despite the fact that there are efforts to increase research in this type of systems. In the Spanish language we have as an outstanding advance [10], a QA system representation of bidirectional transformer encoder that works on the translation of the SQuAD dataset to Spanish by means of the TAR technique.

Therefore, the objective of the research is to design a Model for the Spanish QA problem based on bidirectional transformer encoder representation. As a result of the research, we had an artificial intelligence model that can answer questions formulated in Spanish and English language about the knowledge acquired based on Wikipedia, which is able to answer the doubts that users may have in an open way and not limited to a set of questions and answers.

The present work is divided into the following parts: State of the Art, where the works of the last years regarding QA systems that will serve as a reference point for the design proposed in this research are investigated, Architecture Design, where we will detail the process for the generation of the QA system model in Spanish obtained and Results and discussion.

II. State of the Art

For Question Answering models, recent research has shown a constant use of transformer-based structures.

A. BERT.

BERT uses transformers, an attention mechanism that learns the contextual relationships between words (or subwords) in a text. In its basic form, transformers include two separate mechanisms: an encoder that reads text input and a decoder that produces a prediction for the task- Since the goal of BERT is to generate a language model, only the encoder mechanism is necessary.

Unlike directional models, which read the text input sequentially (left to right or right to left), the encoder [4] Transformer reads the entire sequence of words at once. It is therefore considered bidirectional, although it would be more accurate to say that it is non-directional. This feature allows the model to learn the context of a word based on its entire environment (left and right of the word).

When training language models, there is the challenge of defining a prediction target. Many models predict the next word in a sequence (e.g., "El niño llegó a casa de ___"), a directional approach that inherently limits context learning.
BERT can be used for a wide variety of language tasks, while only adding a small layer to the core model, these applications of BERT can be classification tasks such as sentiment analysis, Question and Answer and entity recognition. Which are established through fine tuning training, where most of the hyperparameters remain the same as in BERT training, and Devlin provides specific guidance in his publication on the hyperparameters that require tuning. The BERT team has used this technique to achieve state of the art results on a wide variety of challenging natural language tasks, detailed in Section 4 of the paper.

B. Knowledge-based models

page as While an enormous amount of information is encoded in the vast amount of text of the web, obviously, information also exists in more structured forms. We use the term 'knowledge-based query' which answers the idea of a language query by mapping it to a query over a structured database [4] Like the text-based paradigm for answering questions, this approach dates back to the early days of natural language processing. In short, the goal of knowledge-based QA systems is to abstract concepts as a tool for answering questions.

Solves the problem of naturally formulated questions and answers, that is, with a with a certain degree of complexity. It also seeks the incorporation of techniques used in information retrieval-based types of QA systems applied to knowledge-based QA for the representation of questions and answers [8]. Additionally, it seeks to develop an architecture for point-to-point QA system with Bi-LTSM networks for question representation and knowledge embedding for answer categorization, obtaining an F-1 indicator of 42.8.

Obtains an accuracy of 0.89 [2], presenting a model capable of answering questions formed in a natural way, that is, in human language and which may lend itself to ambiguities. It is based on the knowledge bases described in RDF, which is considered to present a rather tedious language for information extraction. Therefore, a natural language QA system allowed the common user to access this body of knowledge. The author seeks to overcome the two main challenges of structured queries in a natural language which are sentence linking and query composition.

Another type of approach is the one proposed by [11], the authors seek to improve the performance of a knowledge retrieval model by providing the system with a conceptual reasoning of the queries, this proposing a semantic reasoning algorithm. For this, it uses a knowledge tree of concepts as a basis, which defines the individual and compound conceptual representation of words. On this basis, the questions can be categorized by the concepts and semantic relationships present. The knowledge-based reasoning algorithm iteratively searches for the meaning of the relevant information of each question sentence in its knowledge tree until it finds the final answer that satisfies the query. They obtained an accuracy of 0.6526 for the collected dataset.

Obtaining an accuracy of 0.5759, the model of [12] proposes two attention-based methods as aa replacement to recurrent neural networks such as LSTM for answer selection. Questions from the training dataset that seek the same information, but formulated with lower quality, are ranked. The selective attention mechanism is applied to emphasize the lower quality questions and find a balance in the training process with respect to all objects in the body of knowledge. The self-attention mechanism is applied on the instances by estimating the relationship between the objects in the body of knowledge and the answers, thus calculating the loss of

information in the answer based on the quality of the question itself, for the latter a normalization of answer length is applied based on the appropriate answer parameters.

Seeks to solve the question-answer problem by making use of semantic knowledge corpora [13], he seeks to propose a knowledge-based system that Works point-to-point with the use of neural networks. The author applies attention-based clustering methods for the entity relations module, inner attention sentence and structure attention, this results in knowledge-embedded question vectors that will be processed by two layers of bidirectional LSTM. It also proposes a new attention structure called multilevel target attention (MLTA) to make better use of the hierarchical information held by relations. Obtaining a combined accuracy of 0.8229.

The work of [6] obtains an accuracy of 0.65. The authors present a three-phase QA framework. In the first phase, an algorithm is created for locating the key elements within and contextual relationships within the query using the GloVe word vector representation. In the second phase, the construction of the query graph is taken care of, first separating the critical path from the dependency tree and then generating new candidate graphs of better quality to obtain the desired answer using a proprietary algorithm. Finally, the query answer network is selected using the breadth-first search algorithm within the RDF knowledge network.

Reference [14] seeks to solve the QA problem based on Bi-LSTM-based models for processing candidate sequences and the aggregation model. Using a model with a method based on semantic analysis of the questions asked, taking into consideration the importance of the relationships between the entities found in the question due to the difficulty of representing the expressions in natural language through interpretation templates that are used in some cases.

They obtain an accuracy of 0.83 [15], his research work presents an approach to answer queries formulated in natural language using regular expressions and ontology. The system initially detects the type of query using regex, identifying the interrogative pronouns in the query statement. Then the key objects are identified and matched with the classes and properties of their body of knowledge. To finally generate queries in SPARQL and obtain the final answer when there is a matching with the key data of the question.

Su et al. propose a model based on a dynamic memory network, in a human-like way that can take into account the context or related information of the query. First, the entities of the question are extracted, BiLSTM-CRF [16] is used, subsequently it is updated with all possible entities in the body of knowledge, encoded to a word vector to use a GRU encoding. Then, the memory module iterates over the representations by updating the input and output continuously. Obtaining an accuracy of 0.7941.

Reference [17] propose a Bi-LSTM-CRF based model for entity recognition, which has an accuracy of 0.8294. It is composed of three layers, the first one is in charge of embedding characters in word vectors to pass it to the bidirectional LSTM layer, which processes the sentence to a probability matric to find the answer and the detected entities are labeled according to their type and size. Finally, the matrix is processed by the CRF layer to assign the answer probability to the candidate sentences.

Reference [18] obtains an accuracy of 0.521, their proposal seeks to solve the problem of questions and answers that can serve as a link between people and large bodies of knowledge. Making use of trees for the representation of RDF queries that can obtain results similar to those obtained using deep learning that deny query structuring.

Seeks to solve the problem of relation detection in knowledge-based Question Answering systems. He argues that the representation of existing words in methods using attention mechanism for relationships are very long, because each relation is redundant with others and mostly relationships with two levels of distance are also considered as main. This problem limits the effectiveness of the attention mechanisms since they calculate the value of the vectors based on relationships. Obtaining an accuracy of 0.933 [8].

From the studies carried out by [16], [17] both use the same dataset for their tests, we can see that the semantic relationships of the words increases the precision of QA systems, [16] being the model that sought a greater focus on proper recognition of entities and their relationships, despite using similar Bi-LSTM structures.

From the studies carried out by [13], [14] it is found that the results obtained by having a focus on relationships are comparable to those obtained by using techniques based on neural networks, since both present a very close accuracy.
The use of the LSTM technique is used in the research [8], [12], [16]-[19] for the representation of the entities to take into consideration their context in the question asked and add a weight in their valuation to arrive at the result. So, it is a standard technique in the construction of this type of systems.

C. Information retrieval based models

Reference [1] presents a solution to the problem of complex questions and answers, in this case emphasizing the difficulty of extracting questions that require less superficial analysis than short text answers. Thus, he proposes the summarization of texts that can give answers to these questions proposing a model based on reinforcement learning. A new approach to QA systems is proposed, emphasizing how positive the interaction with the user would be in the learning of these systems, the work done is synthesized in a multi-document summarization proposal based on the query, resulting in an accuracy of 0.1417.

Reference [19] propose a model that combines candidate answer re-ranking methods because it seeks to take into consideration various passages of the body of knowledge without increasing the complexity of the candidate string by considering multiple passages to arrive at the answer, the first method is strength-based answer re-ranking, which ranks answers according to how many times they appear in different text passages. The second method is coverage-based answer re-ranking, which ranks the answer based on whether the union of all retrieved passages covers the most aspects included in the question. For answer extraction we previously used word matching using Bi-LSTM and then re-rank the answers, thus obtaining an F-1 indicator of 0.632.

Increased the final accuracy by using the information from the home services tools in order to expand the question representation [20]. Bi-LSTM is used for word encoding and decoding. For answer generation, the concept of attribution is devised to measure the importance of words in response to the

fact that neural networks can be affected by irrelevant terms in the questions.

Other research [21] seeks to solve the problem of QA systems focused on education, obtaining as a result a model with accuracy 0.763, the proposed model based on BiDAF consists of three modules, using a dynamic concept network that represents entity nodes, edges and inferred knowledge. The query analysis module uses an algorithm that extracts the entities and their respective network of related concepts, then labels the query words and obtains the answer that has the closest approximation to the prefixes found in the query. Finally, the answer extraction module calculates the proximity of the answer based on the similarity of semantic vectors that evaluate the correctness of the entities and relationships.

Reference [20] seeks to propose a model for Question Answering that uses Convolutional Neural Networks as an alternative to the recurrent neural networks used in most QA systems. In addition, the proposed system will seek text processing in the Chinese language, the one obtained from Tencent AI Lab is used as a document corpus. The model measures the similarity of queries with a cosine function with the word vectors of the document corpus. The hinge loss function is used for the convolutional neural network and the adaptive moment estimation (Adam) optimization algorithm is minimized. An accuracy of 0.601 is obtained.

Shao proposes a model that uses a feature extractor followed by three strategies in the matching layer: average weighted clustering, which has an accuracy of 0.564. The initial layer of the transformer is the word representation layer, which processes the text into vectors. [22] Then, the transformer-based extractor uses a self-attention multi-header to model the contextual information of the query, a BiLSTM is used to synthesize the sequential features in the sentence. Finally, the answer is obtained based on the matching strategies.

Propose a new architecture for the model which they call HQACL. This model uses a multi-granular embedding for query representation, which happens at the character, word and contextual word level using the BERT technique. [23] Once the contextual and query presentation vectors are obtained, they are processed through a multi-layer Bi-LSTM network for the high-level representations and their contextualized representation. The proposed model achieves an accuracy of 0.457.

Propose a new attention-based multi-layer aggregated coding transformer (MTA) and an answer generation network based on the created transformer (AG-MTA) in order to make efficient use of contextual information [4] at different levels of text sequences. The structure of the transformer as a network uses a combination of a multi-attention mechanism, which sends the outputs to a sequence of neural networks obtaining an F-1 indicator of 0.803.

They presented a model for Chinese QA systems based on convolutional neural networks and Bi-LSTM to extract the contextual representation of the query [24], this model obtained an accuracy of 0.860. At first, they arrive as text vectors using pooling to obtain the relevant information of the query. Then, an attention mechanism and a co-attention mechanism are combined using an incidence matrix to give the representation of semantic features and interactions.

Reference [25] propose a new method called ASHLK, which consists in pre-processing the data using text segmentation, thus removing words that are not considered meaningful and grouping words that have the same semantic root. Then, given a user's query, the possible answers are retrieved from the document as a set of sentences, the words are scored using a graph representation of all the other candidate answers based on two factors: the relevance to the user's query and the similarity to other candidate answers of high relevance. Finally obtaining with the proposed model an accuracy of 0.720.

From the study [23] it can be said that not always the use of different datasets for the training of the models increases the final accuracy of the system, despite combining BERT with BiLSTM obtaining an accuracy of 0.457.

In the research [24] makes use of the convolutional networks in the same way as [3], the difference is that [24] focuses more on the attention mechanisms for the BiLSTM representation thus obtaining an accuracy of 0.860, while [3] focuses on obtaining an answer through established filter thus obtaining an accuracy of 0.601 for mixed texts.

Reference [25] presents a model focused on body of knowledge processing, he focuses on graph representation which allows him to differentiate the text answer by discerning it from the other candidate strings based on NLP metrics, obtaining an accuracy of 0.720.

It can be evidenced that the contributions in both categories tend to the use of the latest Machine Learning techniques such as Bi-LSTM, RNN, CNN and others, although both categories have a similar structure, the difference is the contribution that the researchers make in the proposal. Thus, knowledge-based models focus more on the representation of the query and body of knowledge to obtain better results, while information retrieval-based models focus on how to extract the answer from the document corpus to obtain better results.

The use of the BERT technique does not always obtain better results, such as the model presented by [23], which obtains an accuracy of 0.457 for a closed domain of a technical nature. However, it is possible to obtain a very high accuracy with proper structuring such as the model presented by [8] that obtains an accuracy of 0.933 outperforming any other research reviewed.

Question Answering models based on other languages obtain results that can be improved, such is the case of the Chinese language whose results are uneven as: [11] with an accuracy of 0.6526, [12] with an accuracy of 0.5759, [16] with an accuracy of 0.7941, [17] with an accuracy of 0.8294, [3] with an accuracy of 0.601 and [24] with an accuracy of 0.860. Additionally, in the Arabic language there is study of [25] that obtains an accuracy of 0.720.

III. Design of the Spanish Question Answering Architecture

To solve the problem of Question Answering in Spanish, a model that takes into consideration the context of the questions asked is presented. The initial question is received by the system, which may or may not have a document corpus from which to extract the answer. In case the user does not include the document corpus with the question, the system

initially seeks to retrieve information related to the question based on the Wikipedia article service. Thus, together with the document corpus and the initial question, the Answer Generation module returns the answer to the initial question. The purpose of the system is to provide answers to the factoid questions posed by users through a web and mobile platform.

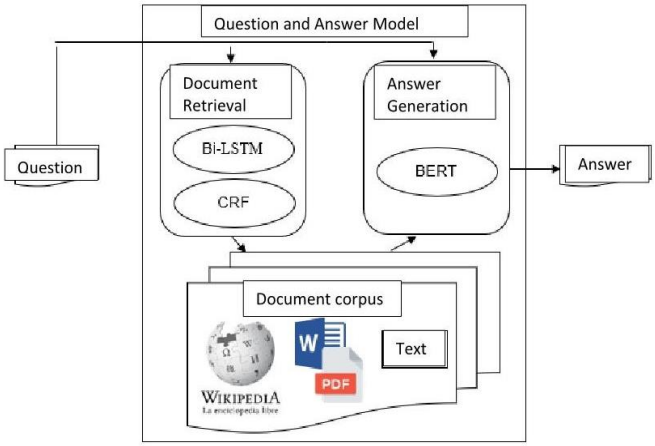

Figure 1. Solution Architecture

A. Dataset

The selected dataset is a translation of the information collection for questions and answers by Standford University, which collects questions and answers from texts of any kind based on a document corpus. Specifically, the version of this dataset that will be used is xQuad, which seeks a translation of this dataset for several languages, including Spanish. SQuAd [23] in its 2.0 version combines 100000 questions from version 1.1 with another 50000 irresponsive questions written contradictorily by the contributors in order to assimilate the questions with those that do have answers. Thus, to obtain a good accuracy in this benchmark, the system must not only know how to obtain an answer, but also identify when to refrain from answering.

Document corpus	Questions
El calor necesario para hervir el agua y suministrar el vapor puede derivarse de varias fuentes, generalmente de **la quema de materiales combustibles** con un suministro adecuado de aire en un espacio cerrado (llamado de varias maneras: **cámara de combustión**, chimenea...). En algunos casos la fuente de calor es un reactor nuclear, energía geotérmica, **energía solar** o calor residual de un motor de combustión interna o proceso industrial. En el caso de modelos o motores de vapor de juguete, la fuente de calor puede ser un calentador **eléctrico**.	1. ¿Cuál es la fuente de calor habitual para hacer hervir el agua en la máquina de vapor?
	2. Aparte de cámara de combustión, ¿que otro nombre que se le da al espacio en el que se quema el material combustible en el motor?
	3. Junto con el calor residual de la energía nuclear, geotérmica y de los motores de combustión interna, ¿qué tipo de energía podría suministrar el calor para una máquina de vapor? ´
	4. ¿Qué tipo de elemento calefactor se utiliza a menudo en las máquinas de vapor de juguete?

Table 1. Comparison of results.

B. Components

1) Question

height This element is the question asked by the user, which is answered by the artificial intelligence mode. In Table 1, the Questions column contains examples of questions e.g. question 1 "¿Cuál es la fuente de calor habitual para hacer hervir el agua en la máquina de vapor?".

2) Document retrieval

This module only works when a query is going to be made to Wikipedia and not when there is already a good document corpus. Document retrieval is in charge of transforming the initial query into a document corpus in the case that the query does not have one, through which we can extract the information that serves to answer the question. This module uses seq2seq techniques using the BiLSTM Encoder-Decoder structure, in charge of transforming the initial query into a representation that in this case will be word vectors and a CRF layer for the contextual relationships of the entities, thus obtaining an adequate query to retrieve the document corpus from the Wikipedia search service. Finally, it returns the Wikipedia article of the topic mentioned in the query to the Answer Generation module.

Initially, the first Encoder layer of the model is in charge of mapping the words of the sentence into the word vector based on word embedding. Thus, the second layer of the model is in charge of extracting the features in the entities present in the query in an iterative manner, since this recurrent neural network architecture is good for dealing with the vanishing gradient problem.

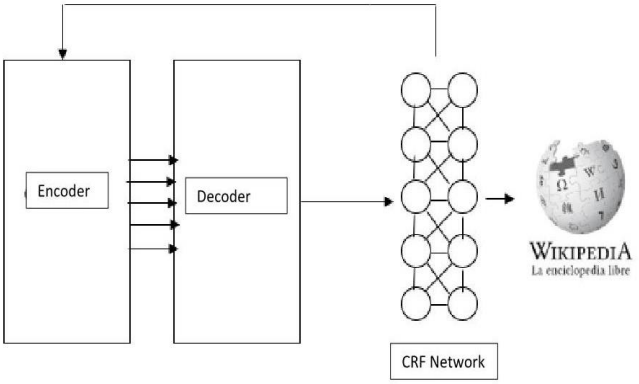

Figure 2. Document Corpus Retrieval Module

Using Word embedding, the representations resulting from the processing have a word vector representation which will be further processed by the CRF network to categorize the contextual enrichment of the query. Having as input the word vector by the BiLSTM, this component is in charge of adding the contextual sense of the initial query, which are recurrently qualified until an evaluation of each entity is achieved based on the BIO categorization, which tells us if an entity (B or I) belongs to another one semantically.

The network of conditional random fields is in charge of labeling each word in the instance, in addition, it maps the dependency relationships between neighboring elements of the query matrix and possible queries to Wikipedia. Iteratively processing the query until each element is parsed. When the final result is obtained, the corresponding Wikipedia search is performed to obtain the document corpus from which the answer to the question is extracted. Once the article related to the topic of the initial question is obtained, the information is sent to the module for generating the answer as text.

3) Document corpus

This element is the textual content from which the question sentence is extracted, it can be a Wikipedia article, a PDF or Word file or even a text provided by the user. It is received in plain text as a BERT input for question generation. In Table 1, the document corpus column contains an example document corpus for the question "¿Cuál es la fuente de calor habitual para hacer hervir el agua en la máquina de vapor?", since the answer is found within the text.

4) Answer generation

The answer generation module is responsible for the selection of the answer, based on the document corpus obtained and the initial query made. It receives both parameters to find the answer to the question that lies within the document corpus; if it is not able to find an answer, this case will be indicated.

Once the initial query and the document corpus obtained are received, BERT is in charge of generating the final answer, since BERT is a model that fulfills the pre-trained tasks of word masking and prediction of the next sentence. The model was trained to perform the necessary Question Answering function using the XQuAD dataset, so that the model has the ability to answer questions related to a given context.

Since BERT is a structure formed by RNN and attention-based methods, it is capable of extracting the relationship within a sentence that words have with respect to each other. The aim is to change the behavior by training BERT with the XQuAD dataset so that, taking advantage of its ability to infer words based on others, it can be able to find the answer to the question that lies "hidden" within the document corpus.

A linear layer of BERT in its Spanish version (BETO) is used. This trained semantic representation model tokenizes the information to be processed in a parallel way. The BERT-based model configuration counts the size of 12 transformers for processing the question with its contextual body. In order for the generation of the answer to be necessary, the input of the BERT layer with the question sentence and the document corpus is properly labeled (in Figure 3, A is the label for the text corresponding to the question and B is the label for the text corresponding to the document corpus).

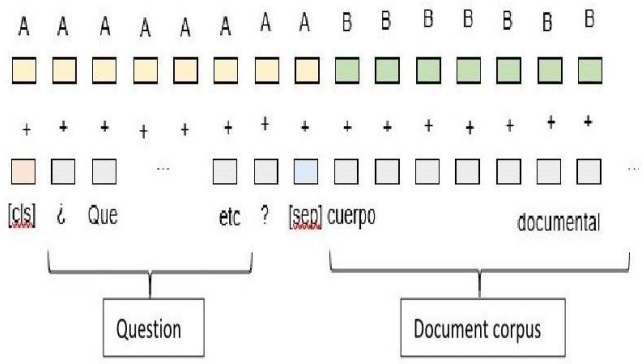

Figure 3. Input Scheme

In Figure 3, each token using BERT is processed in parallel by evaluating whether each token has greater significance of being the initial or final token. The size of the vectors is 768 which are processed sequentially for each token. The purpose of this processing is to identify the tokens within the document corpus that are able to satisfy the initial query, resulting in the tokens that have greater weight for the answer based on the question, being both the initial token as the token where the answer starts and the final token as the token where the answer ends. Therefore, through the Softmax function, we have the probability that a token, which will assign a value between 0 and 1 with respect to the probability of corresponding to the answer sentence.

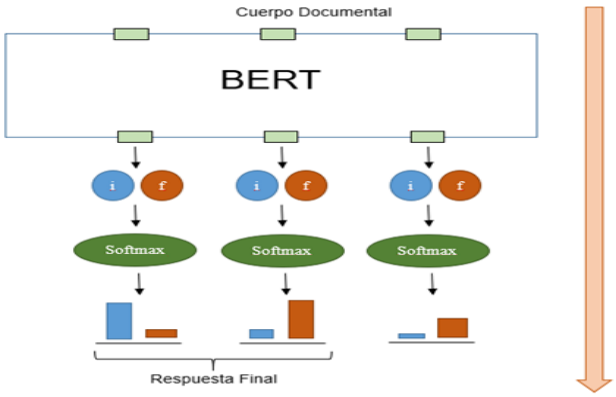

Figure 4. Answer generation

In Figure 4, the document corpus is analyzed based on the question sentence, returning by supervised learning, thanks to the XQuAD dataset, the weight that each token has of being the initial (i) or final (f), applying the Softmax function we can find this probability. Thus, the final answer is obtained by returning the sentence formed within the document corpus starting from the initial token identified to the final token identified.

5) Answer

This element is the textual content that answers the question initially asked, it is obtained as a plain text. In case the question is irresponsive, it will be indicated as such. In Table 1 the document corpus column contains examples of answers highlighted within the document corpus, for example, for question 1 the answer within the document corpus is "the burning of combustible materials".

C. Solution architecture

Following a service-oriented architecture, graphical interfaces to access the platform will be separated for 2 channels: via web browser and via mobile application. Web services were deployed to process user queries, these will be in communication with external services and will be deployed in a Google Cloud container.

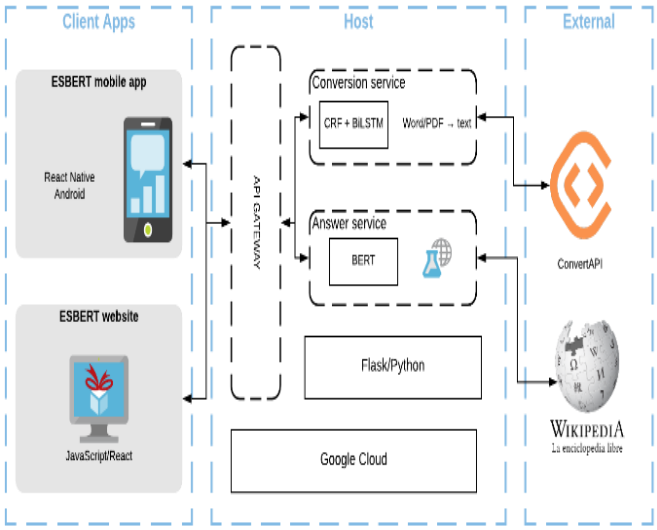

Figure 5. Solution Architecture

The following technologies are used:

Flask: It is self-named as a Python microframework for creating web applications, i.e., dynamic web pages, APIs, etc.

TensorFlow Serving: It is a flexible and high-performance publishing system for machine learning models, designed for production environments. TensorFlow Serving facilitates the implementation of new algorithms and experiments, while maintaining the same server architecture and API.

Google Cloud: A platform that offers more than 90 information technology services in the cloud.

Wikipedia API: It is a web service that allows access to some wiki functionalities such as authentication, page operations and search. It can provide metadata about the wiki and the logged user.

Convert API: It is a web service that allows the conversion of Word, Excel, PowerPoint, HTML, PDF and image files.

IV. Results and Discussion

In this section we present the results of the model based on bidirectional encoder representation transformers in such a way as to expose the results and discussion point obtained from its training with the Spanish version of the SquAD dataset. We then explain how we apply the evaluation metrics for Question Answering models to the results obtained from the model. Thus, from the analysis of these results in comparison to the others, recommendations for further research in this area of knowledge are made.

A. Results

The instances that are present in the dataset have approximately 100'000 questions that are grouped with content from 500 articles extracted from Wikipedia. For the purposes of training and validation of the model based on bidirectional transformer representation, a collection of 10'000 questions from the mentioned articles is counted as validation set. Thus, having a 90/10 split for our presented model.

The training has been performed with the 90'000 data of the training set of the SquAD 1.1 dataset, for which the following configurations have been considered:

Training batch size: 24, the number of instances of the dataset that are loaded in memory at most.

Learning ratio: 3e-5, defines the Adam optimization algorithm.

Number of epochs: 2.0, each of the 90'000 questions in the training set is loaded into the model 2 times.

Do_lower_case: The processing of each text sentence is performed only taking into account its lowercase value.

Max_seq_length: 384, number of tokens that the model can accept in a sentence.

Doc_stride: 128, maximum number of characters that a token can have.

The training was performed in Google Cloud Platform with a cloud instance using the TPU (Tensor Processing Unit) Tesla K80 model, with 12 GB of RAM memory that can deliver up to 1.87Tflops of performance. For the storage of the model, storage buckets were used in Google Storage taking about 6 hours to finish the training making use of the configuration script provided by Google for Question Answering models in BERT, which contains parameterized input and format of the dataset based on the proposed model with some flags that easily allow switching between optimization algorithms and so on. For the validation of the model, we continued using the validation set, which consists of approximately 10 thousand different questions, which were used for training. These answers are presented by our BERT-based model in the following format: id and answer.

Following the format provided by the dataset, the id is a unique identifier that allows us to identify to which instance (question) the answer belongs. This is used for the subsequent calculation of metrics with which our model is compared to other solutions.

No.	Expected Answer	Obtained Answer
1	ímpetu	una fuerza innata de ímpetu
2	Pedro Menéndez de Avilés	Pedro Menéndez de Avilés
	…	
10566	2014	2014

Table 2. Generated Answers

For the processing of the results, certain basic operations were performed that allow us to work on the basis of the results obtained, these operations are applied to each prediction obtained and their comparison with the answer found in the dataset. These operations were the following: The elimination of articles in order to count only the meaningful tokens, those that carry a syntactic value for the sentence. The elimination of punctuation marks, in order to take into account only the words. Transforming each letter of the sentence to lower case, for better matching. Separating the tokens of a sentence for individual counting.

Given the nature of the questions as not being uniquely correct, the number of correct tokens in the answer sentence was taken as the value to which to attribute the success of the model. Thus, except for the items, we calculated the total number of correctly obtained tokens over the total number of tokens obtained. Based on these considerations the following metrics are obtained:

Metric	Value

Accuracy	78.5537
F-1	76.5025
EM	61.6076

Table 3. Obtained Model Metrics

B. Discussion

With the results obtained from our BERT-based model, a comparison was made between the other works that used SQuAD dataset. Table 4 shows the comparisons between the investigations and our proposal. It is worth mentioning since our model obtains an accuracy of 0.7856 it should be considered successful since the values are very close to those of the state of the art that mostly their values oscillate between those obtained from the studies for both English and Chinese language.

Author	Accuracy	F-1
Our proposal	0.7856	0.765
Liu Chen [26]		0.428
Sen Hu [2]	0.89	0.78
Ziqi Lin [11]	0.6526	
Mengxi Wei [12]	0.5759	
Run-Ze Wang [13]	0.8229	
Hai Jin [6]	0.65	0.40
Yunshi Lan [14]	0.809	
Maria Rahim [15]	0.83	
Lei Su [16]	0.7941	
ChuanLong Li [17]	0.8294	
Shuguang Zhu [18]	0.521	
Yongrui Chen [8]	0.933	
Yllias Chali [1]	0.0878	
Shuohang Wang [19]		0.632
Mengyang Zhang [20]		0.476
Anant Agarwal [21]	0.763	
Taihua Shao [22]	0.564	
Yongping Du [23]	0.457	
Shengjie Shang [4]		0.803
Lin-Qin Cai [24]	0.860	0.866
Asad Abdi [25]	0.72	0.604

Table 4. QA Models Results

Table IV shows a compilation of the values obtained from each study for the accuracy metrics, F-1 and Exact Match. The results obtained by the present study being in the upper half with the highest accuracy of all, we can say that the result was satisfactory. However, the precision obtained for our model based on BERT in Spanish with 0.7856 accuracy, comparing with [8], which is also based on BERT, reaches 0.933 accuracy, in this case we present the model adapted for Spanish.

The effectiveness of the use of techniques based on transformers applied to Question Answering of different languages is also evidenced by the studies of [16], [17], [8] and ours with accuracies of 0.7941, 0.8294, 0.933, and 0.7856 respectively. In comparison with the accuracy obtained by studies such as [16], [22] with 0.7941 and 0.564 respectively apply techniques based on graphs and word vectors.

The model obtained based on the BERT model for Spanish BETO, presents a better performance for a foreign language as opposed to another non-English language model based on the multilingual version. Thus, suggesting that the advances of this NLP structure should be directed to provide a solid base for each language and thus improve the obtained results. The proposed final system can answer the questions with a probability close to 80% so that 1 out of 5 questions may not be correctly answered. This leaves ample room for improvement so that the question-and-answer platform to support school students can be highly reliable.

Conclusions

The result improved by entering QA to a more specialized BERT model adapted for the Spanish language, the multilingual version of BERT and the Spanish version of BETO were used. The study of QA systems in Spanish could be diversified if a dataset of questions and answers oriented only to Spanish is generated, since certain contextual limitations and idioms to which translated datasets are subject could be avoided.

References

[1] Y. Chali, S. A. Hasan, and M. Mojahid, “A reinforcement learning formulation to the complex question answering problem,” Inf. Process. Manag., vol. 51, no. 3, pp. 252–272, 2015, doi: 10.1016/j.ipm.2015.01.002.

[2] S. Hu, L. Zou, J. X. Yu, H. Wang, and D. Zhao, “Answering Natural Language Questions by Subgraph Matching over Knowledge Graphs,” IEEE Trans. Knowl. Data Eng., vol. 30, no. 5, pp. 824–837, 2018, doi: 10.1109/TKDE.2017.2766634.

[3] G. Zhang, X. Fan, C. Jin, and M. Wu, “Open-domain document-based automatic QA models based on CNN and attention mechanism,” Proc. - 10th IEEE Int. Conf. Big Knowledge, ICBK 2019, pp. 326–332, 2019, doi: 10.1109/ICBK.2019.00051.

[4] S. Shang, J. Liu, and Y. Yang, “Multi-Layer Transformer Aggregation Encoder for Answer Generation,” IEEE Access, vol. 8, pp. 90410–90419, 2020, doi: 10.1109/ACCESS.2020.2993875.

[5] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuad: 100,000+ questions for machine comprehension of text,” EMNLP 2016 - Conf. Empir. Methods Nat. Lang. Process. Proc., pp. 2383–2392, 2016, doi: 10.18653/v1/d16-1264.

[6] H. Jin, Y. Luo, C. Gao, X. Tang, and P. Yuan, “ComQA: Question Answering Over Knowledge Base via Semantic Matching,” IEEE Access, vol. 7, pp. 75235–75246, 2019, doi: 10.1109/ACCESS.2019.2918675

[7] C. E. Serquen-Llallire, H. D. Calderon-Vilca, and F. C. Cardenas-Marino, “Comparison of two Algorithms for Routing Questions and Answers, Applied to Group of Students Software Engineering,” 2018 IEEE Lat. Am. Conf. Comput. Intell. LA-CCI 2018, 2019, doi: 10.1109/LA-CCI.2018.8625262.

[8] Y. Chen and H. Li, “DAM: Transformer-based relation detection for Question Answering over Knowledge Base,” Knowledge-Based Syst., vol. 201–202, p. 106077, 2020, doi: 10.1016/j.knosys.2020.106077.

[9] UNICEF, “Education and COVID-19,” 2020. https://data.unicef.org/topic/education/covid-19/.
[10] C. P. Carrino, M. R. Costa-Jussà, and J. A. R. Fonollosa, “Automatic Spanish translation of the SQuAD dataset for multilingual question answering,” Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc., no. May, pp. 5515–5523, 2020.
[11] Z. Lin, W. Ni, H. Zhang, M. Zhao, and Y. Liu, “Semantic reasoning of question answering over heroes of the marshes based on concept knowledge tree,” Proc. - 2017 10th Int. Symp. Comput. Intell. Des. Isc. 2017, vol. 2, pp. 244–247, 2018, doi: 10.1109/ISCID.2017.36.
[12] M. Wei and Y. Zhang, “Natural Answer Generation with Attention over Instances,” IEEE Access, vol. 7, no. c, pp. 61008–61017, 2019, doi: 10.1109/ACCESS.2019.2904337.
[13] R. Z. Wang, Z. H. Ling, and Y. Hu, “Knowledge Base Question Answering with Attentive Pooling for Question Representation,” IEEE Access, vol. 7, pp. 46773–46784, 2019, doi: 10.1109/ACCESS.2019.2909826.
[14] Y. Lan, S. Wang, and J. Jiang, “Knowledge Base Question Answering with a Matching-Aggregation Model and Question-Specific Contextual Relations,” IEEE/ACM Trans. Audio Speech Lang. Process., vol. 27, no. 10, pp. 1629–1638, 2019, doi: 10.1109/TASLP.2019.2926125.
[15] M. Rahim, Z. Turabee, Q. Rajput, and S. A. Khoja, “Semantic Based Question Answering System on Travel Ontology,” 2019 6th Int. Conf. Soc. Networks Anal. Manag. Secur. SNAMS 2019, pp. 67–74, 2019, doi: 10.1109/SNAMS.2019.8931886.
[16] L. Su, T. He, Z. Fan, Y. Zhang, and M. Guizani, “Answer Acquisition for Knowledge Base Question Answering Systems Based on Dynamic Memory Network,” IEEE Access, vol. 7, pp. 161329–161339, 2019, doi: 10.1109/ACCESS.2019.2949993.
[17] C. L. Li, H. X. Liu, F. R. Zhang, and Y. Q. Feng, “Research on entity recognition method in knowledge base question answering,” Proc. - 2019 18th Int. Symp. Distrib. Comput. Appl. Bus. Eng. Sci. DCABES 2019, pp. 128–131, 2019, doi: 10.1109/DCABES48411.2019.00039.
[18] S. Zhu, X. Cheng, and S. Su, “Knowledge-based question answering by tree-to-sequence learning,” Neurocomputing, vol. 372, pp. 64–72, 2020, doi: 10.1016/j.neucom.2019.09.003.
[19] S. Wang et al., “Evidence aggregation for answer re-ranking in open-domain question answering,” 6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc., vol. 1, pp. 1–14, 2018.
[20] M. Zhang, G. Tian, and Y. Zhang, “A Home Service-Oriented Question Answering System with High Accuracy and Stability,” IEEE Access, vol. 7, no. c, pp. 22988–22999, 2019, doi: 10.1109/ACCESS.2019.2894438.
[21] A. M. A. Agarwal, N. Sachdeva, R. K. Yadav, V. Udandarao, V. Mittal, A. Gupta, “Eduqa : Educational Domain Question Answering System Using,” ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., pp. 8137–8141, 2019.
[22] T. Shao, Y. Guo, H. Chen, and Z. Hao, “Transformer-Based Neural Network for Answer Selection in Question Answering,” IEEE Access, vol. 7, pp. 26146–26156, 2019, doi: 10.1109/ACCESS.2019.2900753.
[23] Y. Du, W. Guo, and Y. Zhao, “Hierarchical Question-Aware Context Learning with Augmented Data for Biomedical Question Answering,” Proc. - 2019 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2019, pp. 370–375, 2019, doi: 10.1109/BIBM47256.2019.8983185.
[24] L. Q. Cai, M. Wei, S. T. Zhou, and X. Yan, “Intelligent Question Answering in Restricted Domains Using Deep Learning and Question Pair Matching,” IEEE Access, vol. 8, pp. 32922–32934, 2020, doi: 10.1109/ACCESS.2020.2973728.
[25] A. Abdi, S. Hasan, M. Arshi, S. M. Shamsuddin, and N. Idris, “A question answering system in hadith using linguistic knowledge,” Comput. Speech Lang., vol. 60, 2020, doi: 10.1016/j.csl.2019.101023.
[26] L. Chen, G. Zeng, Q. Zhang, X. Chen, and D. Wu, “Question answering over knowledgebase with attention-based LSTM networks and knowledge embeddings,” Proc. 2017 IEEE 16th Int. Conf. Cogn. Informatics Cogn. Comput. ICCI*CC 2017, no. 2015, pp. 243–246, 2017, doi: 10.1109/ICCI-CC.2017.8109757.

Author Biographies

Robert C. Gutierrez Ramos Bachelor’s degree in software engineering, passionate about Artificial Intelligence, mathematics and Natural Language Processing. Software Architect for banking companies. He is currently looking to open new projects on GPT-3 based models and participate in international research collaboration.

Hugo D. Calderon Vilca PhD in Computer Science, research professor of the "Artificial Intelligence" Group of the Universidad Nacional Mayor de San Marcos - Peru, advisor of undergraduate and graduate thesis projects related to Neural Networks, Machine Learning and Natural Language Processing. Professor of doctoral programs in other universities.

Flor C. Cardenas Mariño PhD in Computer Science, research professor of the Universidad Nacional Mayor de San Marcos, member of the Artificial Intelligence Research Group, member of the OPTIMACO Research group. Experience in research projects funded by CONCYTEC using optimization algorithms in the field of Artificial Intelligence.