Evaluating Recognizing Question Entailment Methods for a Portuguese Community Question-Answering System about Diabetes Mellitus

Thiago Castro Ferreira[1] **João Victor de Pinho Costa**[1] **Isabela Rigotto**[1] **Vitoria Portella**[1]
Gabriel Frota[1] **Ana Luisa A. R. Guimarães**[1] **Adalberto Penna**[1] **Isabela Lee**[1] **Tayane A. Soares**[1]
Sophia Rolim[1] **Rossana Cunha**[1] **Celso França**[1] **Ariel Santos**[1] **Rivaney F. Oliveira**[1] **Abisague Langbehn**[2]
Daniel Hasan Dalip[3] **Marcos André Gonçalves**[1] **Rodrigo Bastos Fóscolo**[1] and **Adriana Pagano**[1]
[1]Federal University of Minas Gerais (UFMG), Brazil
[2]Technical High School of UFMG, Brazil
[3]Federal Center for Technological Education of Minas Gerais (CEFET-MG), Brazil
thiagocf05@ufmg.br

Abstract

This study describes the development of a Portuguese Community-Question Answering benchmark in the domain of Diabetes Mellitus using a Recognizing Question Entailment (RQE) approach. Given a premise question, RQE aims to retrieve semantically similar, already answered, archived questions. We build a new Portuguese benchmark corpus with 785 pairs between premise questions and archived answered questions marked with relevance judgments by medical experts. Based on the benchmark corpus, we leveraged and evaluated several RQE approaches ranging from traditional information retrieval methods to novel large pre-trained language models and ensemble techniques using learn-to-rank approaches. Our experimental results show that a supervised transformer-based method trained with multiple languages and for multiple tasks (MUSE) outperforms the alternatives. Our results also show that ensembles of methods (stacking) as well as a traditional (light) information retrieval method (BM25) can produce competitive results. Finally, among the tested strategies, those that exploit only the question (not the answer), provide the best effectiveness-efficiency trade-off. Code is publicly available[1].

1 Introduction

Question answering (QA) aims to automatically retrieve precise, rather than merely relevant, answers to a given question. The field has faced exponential progress along the years with new corpora (Rajpurkar et al., 2016; Ahmad et al., 2019) as well as computational models which approach the task from different perspectives. One of these is known as Recognizing Question Entailment (RQE).

Given a premise question (aka *query*), a RQE approach aims to retrieve semantically similar archived questions which have been already answered (Ben Abacha and Demner-Fushman, 2019). The task became relevant thanks to popular Community QA forums, such as Yahoo Answers, Quora and Stack Overflow, where an RQE approach is used to automatically search a large body of material looking up for archived question-answer pairs entailing a user posed question.

Besides popular forums, a domain in which RQE approaches are highly beneficial and have been extensively studied is the medical one. Russell-Rose and Chamberlain (2017) showed that, when using traditional information retrieval search engines to query medical information, healthcare professionals spend on average 60 minutes to formulate a search strategy, 3 minutes to analyse the relevance of a retrieved document, and 4 hours of total search time. Ben Abacha and Demner-Fushman (2019) suggest that healthcare consumers may also benefit from QA systems through which they can ask for desired information in natural language instead of having to perform complex search strategies. In fact, a study reveals that, in 2013, 59% of U.S. adults searched for health information online and 35% used healthcare search engines to figure out what medical condition they or someone else had[2].

To advance the state-of-the-art in RQE approaches for medical-specific applications, some benchmarks have been proposed. Targeting Frequently Asked Question (FAQ) by healthcare consumers, Abacha and Demner-Fushman (2016) introduced a collection of 8,890 pairs of questions labelled as having or not the same meaning. In the TREC 2017 LiveQA track, a medical question answering task was proposed addressing the automatic answering of consumer health questions received by the U.S. National Library of Medicine (Abacha et al., 2017). Under the shared-task, a

[1]https://github.com/Dia-Bete/RANLP2021

[2]https://www.pewresearch.org/internet/2013/01/15/health-online-2013/

Proceedings of Recent Advances in Natural Language Processing, pages 234–243
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_028

total of 738 question-answer pairs were publicly released. Finally, Ben Abacha and Demner-Fushman (2019) released MedQuad, a dataset with 47,457 medical question-answer pairs of 37 question types extracted from 12 websites of the the National Institutes of Health of United States.

Although the RQE task has been extensively studied, as is the case with other NLP subfields, studies and approaches largely focus on the English language. In this study, we goes in a different direction, facing the challenge of developing an NLP approach for a low resource language. Specifically, our study focuses on the Portuguese language and investigates the development of a Community Question Answering system in the particular domain of Diabetes Mellitus using RQE. This allows investigating new challenges in terms of multi-lingual NLP problems.

From an application domain (Health) perspective, world-widely, the number of people with diabetes increased from 108 million in 1980 to 422 million in 2014[3]. It is expected that the disease will reach 629 million in 2045. These numbers are partially explained by the rapidly increase of the disease in low- and middle-income countries, such as Brazil, where around 12 million people have this health condition. Aiming to cope with this disease and to have a better quality of life, a significant number of Portuguese speakers with this condition engages in dedicated public forums. In order to improve the access to information about this health condition for these speakers, this study aims to leverage RQE approaches to build BeteQA, a Portuguese community question-answering system to provide prompt and accurate answers to questions about Diabetes Mellitus posed by social forum users.

To meet our goal, similar to Abacha and Demner-Fushman (2016); Abacha et al. (2017) and Ben Abacha and Demner-Fushman (2019), we first built a Portuguese benchmark with 785 pairs between premise questions and archived answered questions annotated as perfect match, relevant and irrelevant, following Nakov et al. (2016, 2017). Based on our benchmark corpus, we leveraged and evaluated several RQE approaches ranging from traditional information retrieval (IR) methods (BM25 and TF-IDF cosine similarity) to novel large pre-trained language models such as BERT (Devlin et al., 2019). For the latter methods, we evaluated them according to both zero-shot and fine-tuned learning setups. We also used ensemble models (stacking) based on a learn-to-rank model to combine the outputs of the previous methods with other traditional linguistic features.

Experimental results show that a supervised transformer-based method trained in multiple languages and for multiple tasks (MUSE) outperforms the alternatives in a zero-shot setting. Moreover, results show that the ensemble method (stacking) as well as a traditional (light) IR method (BM25) have the potential to provide competitive results. Finally, among the tested strategies, those that exploit only the question (not the answer), provide the best effectiveness-efficiency trade-off. Our failure case analysis also reveals that most failures occur for longer sentences containing a smaller number of relevant candidates and harder separability.

The remainder of this paper is organized as follows: Section 2 describes how the benchmark was built. Section 3 explains the RQE models leveraged to rank similar questions about Diabetes. Section 4 describes the experimental methods while Section 5 discusses the experiment results and a failure analysis. Section 6 presents related work and Section 7 concludes our study.

2 Data

Collection To build a Portuguese RQE benchmark in the domain of Diabetes Mellitus, we first manually extracted Portuguese questions about this health condition from specialized Websites and Social Media forums. In particular, most questions were extracted from the FAQ section in the website of the Brazilian Association about Diabetes[4] as well as in forums about this health condition, such as *DIABETES - DIABÉTICOS*[5], a Facebook community about Diabetes with around 85 thousand Portuguese speaking users.

Preprocessing and Anonymization To keep users' privacy, the extracted questions from forums were manually de-identified by first removing emojis, fixing orthographic mistakes and paraphrasing non-fluent syntactic structures which could point to idiosyncratic wordings by users. Then any identifier, such as name, phone or address, was removed from the questions. Moreover, *quasi-identifiers*,

[3] https://www.who.int/news-room/fact-sheets/detail/diabetes

[4] https://www.diabetes.org.br/

[5] https://www.facebook.com/groups/298949446842231/

such as age and relative mentions, were modified. Users' age were modified by randomly choosing a number in the interval of $[age-5; age+5]$, whereas mentions to relatives were randomly changed by the reference to a relative with similar age, such as *parent* $\leftrightarrow$ *uncle, parent in law*; *sibling* $\leftrightarrow$ *cousin, partner*; *soon* $\leftrightarrow$ *nephew*. We collected a total of 1474 questions.

Answers 4 Medical students were recruited to review the question-answer pairs extracted from FAQ sections of websites as well as to formulate answers to the de-identified questions from public online forums. During the evaluation process, each question was answered by a single medical student. In case of doubts, the respective question would be discussed among the students together with a medical specialist. To organize the process, the students kept a standardized database of answers to the collected questions. Each answer in this database was classified according to 10 topics: 1) General information about Diabetes; 2) Diagnosis; 3) Chronic complications; 4) Acute complications; 5) Treatment; 6) Treatment control; 7) Comorbidities; 8) Signs and symptoms; 9) Motivation; and 10) Highly frequent, though unrelated to diabetes.

Answers were elaborated pursuant to Article 37 of Chapter V of the Brazilian Code of Medical Ethics[6], which prohibits treatment prescription without actual patient examination. Hence, the answers were constructed with the aim of informing the user about diabetes and related issues, without offering any diagnosis or treatment. In cases where the user requested some type of intervention, answers were prepared in order to guide them to seek a public healthcare unit, both to obtain an accurate diagnosis and to have adequate therapeutic plans designed by healthcare professionals.

RQE benchmark To finally build the benchmark, we randomly selected 200 questions (roughly 15%) as premises, whereas the remaining 1274 (together with their answers) were indexed by a BM25 model (Jones et al., 2000). For each premise question, we retrieved the 5 most similar candidate questions, together with their answers, using BM25. Finally, following (Nakov et al., 2016, 2017), given 1000 triples (premise question, candidate question, candidate answer), a medical student was recruited to annotate whether the candidate question was a perfect match, relevant or irrelevant to the premise one. The candidate question was considered a perfect match when it conveyed exactly the same semantic meaning as the premise question. When both candidate and premise questions shared the same topic, but were not semantically identical, the candidate was labeled as relevant. Otherwise, the candidate question was labeled as irrelevant to the premise question.

Once the annotation was concluded, premise questions with only irrelevant candidate questions were ruled out, resulting in a corpus with 157 premise questions, each one aligned with 5 annotated question-answer pairs.

3 Models

Drawing on our collected Portuguese benchmark about Diabetes, we evaluated several approaches to rank question-answer pairs to their premise questions. Such approaches range from traditional bag-of-words information retrieval techniques to novel methods based on continuous vector representations and Learn-to-rank ensembles.

3.1 Token-Based Approaches

BM25 is a fast information retrieval technique (Jones et al., 2000) which, in the context of RQE, calculates the relevance of archived question-answer pairs to a given premise question using a family of scoring functions based on bag-of-words.

Cosine Similarity over TF-IDF ranks the similarity of archived documents, such as questions or question-answer pairs, to the premise question by computing the cosine similarity between their TF-IDF vector representations (Salton and McGill, 1986). TF-IDF is a bag-of-words technique, standing for *term frequency–inverse document frequency*. As the name implies, a TF-IDF vector representation of a document is computed by counting the frequency of its tokens as well as their specificity, defined by an inverse function of the number of documents in which each of its tokens occurs.

3.2 Embedding-Based Approaches

Currently, sparse bag-of-words vector representations have given place to dense vectors computed by neural networks and popularly known as (word, sentence or document) *embeddings* (Mikolov et al., 2013). We leveraged some of these representations as RQE approaches.

[6] https://portal.cfm.org.br/images/PDF/cem2019.pdf

3.2.1 Skip-Gram Wang2Vec

We used the Brazilian Portuguese word embeddings of 300 dimensions computed by a skip-gram Wang2Vec architecture described in Hartmann et al. (2017). To obtain the embedding representations of a multi-word document such as a question, we first looked up for the word-embeddings of each of its tokens and then averaged them. Absent tokens in the skip-gram Wang2Vec vocabulary were represented by the embedding of the `OOV` (out-of-vocabulary) token. During ranking, we used cosine similarity to measure the semantic distance between a premise question and its candidates.

Limitation Skip-Gram Wang2Vec provides context-free representations of words, i.e. the approach does not distinguish the meaning of a particular occurrence of a word taking its surrounding context into account. For instance, the word *bank* would be represented by the same vector representation in the expressions "financial *bank*" and "river *bank*".

3.2.2 BERT-Based Approaches

More recently, several studies have proposed large neural language models which compute *context-sensitive* word embeddings representations (Howard and Ruder, 2018; Peters et al., 2018; Devlin et al., 2019). These language models take the local word context into account to generate its meaning representation. One of the first and most popular approaches of this kind is the "Bidirectional Encoder Representations from Transformers" (BERT) (Devlin et al., 2019), which encodes the meaning of a word into a vector taking its surrounding words (before and after) into account. BERT is pretrained in an unsupervised fashion using objective functions like word-denoising (i.e., predicting a masked word in a text) and next sentence prediction. In our study, we leveraged some of its multilingual and Brazilian Portuguese variations to the RQE task in the Diabetes Mellitus domain.

mBERT is a BERT multilingual version trained with texts from the top 100 largest Wikipedias languages. We used its base and cased configuration (`bert-base-multilingual-cased`). Given a multi-word document to be encoded, this approach computes the context-sensitive vector representations of its words and averaged them to obtain the document embedding. As done with the skip-grams, during ranking we measured the cosine similarity between the vector representations of a premise question and its question-answer candidate pairs.

BERTimbau is a variation of BERT pretrained only with Brazilian Portuguese texts (Souza et al., 2020). The pretraining process was done with the largest corpus for the language, know as brWaC corpus (Wagner Filho et al., 2018). In our study, we used its large and cased configuration (`bert-large-portuguese-cased`).

BioBERTpt is another variation of BERT which was pre-trained with Brazilian Portuguese Clinical texts (Schneider et al., 2020) for the task of named entity recognition in the target domain.

3.2.3 MUSE

Differently from BERT that trains large language models in an unsupervised fashion style, other studies have sought to learn a semantic vector space among words and documents by training a neural network in a supervised context with multiple downstream tasks. A state-of-the-art approach in this genre is the "Multilingual Universal Sentence Encoder" (MUSE) (Yang et al., 2020). MUSE embeds texts in 16 languages, which are later fed into classification heads for the target task. In particular, the approach is trained for 3 types of tasks: semantic retrieval, bitext retrieval and question answering retrieval. The study also showed that the model can learn cross-lingual vector representations, i.e. it is possible to measure the semantic similarity between two texts of different languages. Despite this cross-lingual resource, we only used the model to embed Portuguese texts about Diabetes Mellitus. In particular, we used the MUSE approach based on the Transformer architecture (Vaswani et al., 2017).

Different from BERT approaches, we followed Yang et al. (2020) and measured the similarity between the vector representations of premise questions and their question-answer candidate pairs using dot product instead of the cosine similarity.

3.3 Learn-to-rank Ensemble (Stacking)

We sought to investigate whether ensembling (i.e., stacking) the semantic similarity measures computed by the previous RQE approaches could boost the results for ranking candidate question-answer pairs to their corresponding premise questions. To fulfill this goal, we trained a Coordinate Ascent

learn-to-rank approach (Metzler and Croft, 2007)[7]. Besides feeding the learn-to-rank approach with semantic features, we also wanted to analyse whether features which assess the quality of the questions can help the RQE task. Relying on Dalip et al. (2012), we used three types of textual quality assessment features as explained next.

3.3.1 Features

Our learn-to-rank approach uses four types of features, one based on the semantic measures computed by the previously described approaches and 3 other which assess the quality of texts based on its length, style and readability. The quality assessment features were computed for both premise questions and their corresponding question-answer candidate pairs.

Semantic Features the semantic similarities computed by our token-based, skip-gram wang2vec, BERT-based, MUSE approaches are used as semantic features by our learn-to-rank ensemble approach.

Length Features we compute several quality assessment features based on length such as the size of a text based on the number of phrases, words and characters.

Style Features we count the total number of phrases higher/lower than the average phrase length in the text; the size of the largest and shorted phrases; the number of articles; prepositions; auxiliary and total number of verbs; coordination, subordinating and correlative conjunctions; indefinite, interrogative, relative and total number of pronouns; an sentences starting with articles, prepositions, auxiliary verbs, general verbs, (coordination, correlative and subordinating) conjunctions and (indefinite, interrogative and relative) pronouns.

Readability Features we computed ARI, Coleman-Liau, Flesch Reading Ease, Flesch Kincaid, Gunning Fog Index, Lasbarhets index and SMOG Grading.

3.3.2 Settings

We trained our Coordinate Ascent learn-to-rank approach with 5 random restarts, 25 iterations to search in each dimension, 0.001 of tolerance and normalizing input features using z-score. The training data for the ensemble of methods was produced with nested cross-validation in the training set, thus avoiding any potential risk of data leakage problems from test to training.

[7]We tested several L2R methods in preliminary experiments including Random Forests and LambdaMART. There were no statistically significant differences among them. As Coordinate Ascent was much faster than the other ones at training time, we chose it due to the many tests we wanted to perform in our experiments.

3.4 Fine-tuned Approach

All the embedding approaches described in Section 3.2 and MUSE (Section 3.2.3) were used in a zero-shot learning setup, i.e. they were not trained specifically to our domain and, in the case of the embedding-based approaches, nor even to a semantic retrieval task such as RQE.

In order to have a fine-tuned embedding-based approach in our analysis, we used a neural classifier based on BERTimbau (in its large and cased version) as a Portuguese RQE approach about Diabetes. Given a premise question and a candidate question-answer pair, BERTimbau works by first encoding both documents. Following other fine-tuning studies with BERT (Devlin et al., 2019), for each document, we chose its embedding representation based on its special token `[CLS]`. Finally, we fine-tune the model by computing the cosine embedding loss function as in Equation 1:

$$loss(x, y) = \begin{cases} 1 - cos(x_1, x_2), \text{if } y = 1 \\ max(0, cos(x_1, x_2)), \text{if } y = 0 \end{cases} \quad (1)$$

where x_1 and x_2 are the `[CLS]` vector representations of the premise question and the candidate question-answer pair, and y is the gold-standard label indicating whether they are similar or not. We treat the problem as binary, merging perfect match and relevant cases of our benchmark as positive instances, whereas the irrelevant ones as negatives. During training, the model backpropagates the gradients of the neural network using the AdamW optimizer with learning rate of 1e-5 and batch size of 4.

4 Evaluation

We evaluated the proposed approaches as a ranking problem. Given a premise question from the described benchmark, our goal is to investigate which model can better rank "Perfect Match" and "Relevant" candidate question-answer pairs ahead of "Irrelevant" ones. Following Nakov et al. (2016, 2017), we treated the problem as a binary one, not distinguishing "Perfect Match" and "Relevant" candidate questions.

	Question-Question			Question-Answer			Question-Question+Answer		
	Rank	**MAP**	**MRR**	**Rank**	**MAP**	**MRR**	**Rank**	**MAP**	**MRR**
BM25	#2	84.83	86.94	**#1**	**87.36**	**89.68**	**#1**	**87.36**	**89.68**
TF-IDF Cosine	#2	84.11	86.77	#3	83.66	85.67	#3	83.51	85.26
Wang2Vec	#4	78.88	81.05	#4	79.33	80.99	#4	79.70	81.52
BERTimbau	#2	84.98	88.00	#3	82.17	84.07	#2	83.86	86.80
mBERT	#4	81.15	83.34	#6	74.06	75.31	#5	77.03	79.78
BioBERT_pt	#4	78.14	81.52	#5	74.48	73.70	#5	78.13	80.33
MUSE	**#1**	**88.59**	**91.58**	**#1**	**87.24**	**88.90**	**#1**	**88.90**	**91.26**
Fine-tuned	**#1**	**87.12**	**90.66**	#3	82.91	83.86	#2	83.92	86.31
L2R Semantic	**#1**	**87.46**	**90.44**	**#1**	**86.88**	**89.21**	**#1**	**88.87**	**91.02**
L2R Quality	#6	71.95	72.85	#5	74.09	74.48	#6	72.35	73.78
L2R All	#2	85.77	89.07	**#1**	**85.82**	**88.31**	#2	85.08	86.60

Table 1: MAP@5 and MRR@5 results of the approaches measuring the similarity among premise questions and candidates through Question-Question, Question-Answer and Question-Question+Answer. Ranking was computed based on pair-wise comparisons among the MAP@5 models with the Wilcoxon Signed-Rank test. Best results are in **bold**, including statistical ties.

Metrics We evaluated the approaches using two popular ranking measures: the Mean Averaged-Precision (MAP) as the main metric and the Mean Reciprocal Rank (MRR) as the secondary one. Since each premise question of our benchmark is attached to 5 candidates, we used this length to compute the metrics (e.g., MAP@5 and MRR@5).

Comparing Strategies The task of Recognizing Question Entailment (RQE) traditionally works by measuring the similarity between two questions. However, in the Semeval task 3 shared-task about Community Question-Answering (Nakov et al., 2016, 2017), some of the leading approaches worked by measuring the similarity of a premise question taking into account both the candidate question and the answer. Moreover, approaches such as Yang et al. (2020) performs the Question-Answering task as a "Recognizing Answer Entailment" style, where the representation of a premise question is directly compared to the representations of the candidate answers. In this study we investigate the three comparing strategies: Question-Question, Question-Answer and Question-Question+Answer.

Cross-validation The approaches were evaluated using cross-validation using 5 folds. The obtained results were averaged across the folds and statistically tested according to the Wilcoxon Signed-Rank test in MAP@5.

5 Results

Overall Analysis: Ranking of Methods Table 1 displays the MAP@5 and MRR@5 results of the approaches in the Question-Question, Question-Answer and Question-Question+Answer strategies. Best results are marked in **bold**, including statistical ties. Regarding the proposed approaches, results show the advantage of MUSE, being the only method together with the ensemble with semantic features (e.g. `L2R Semantic`) to rank first in the three strategies according to the MAP@5. Although there is a tie between both methods, MUSE is a single model, whereas the latter is an ensemble of all our semantic similarity, demanding much more computational resources. For this reason, we assume MUSE as the model with the best results in our benchmark. Interestingly, MUSE was applied to the problem of Portuguese QA about Diabetes Mellitus in a zero-shot learning setup, i.e. it was not optimized to the task and, even so, was ranked first in all the strategies.

Traditional Token-Based Methods the "old-school" BM25 had very competitive results. The approach ranked first in the Question-Answer and Question-Question+Answer strategies and second in the Question-Question one. Besides effective, this approach has the advantage of not demanding a high volume of computational resources.

Word Embeddings and BERT Traditional context-free word embeddings, represented by Wang2vec, did not have a good performance in the evaluation, being outperformed by traditional methods such as BM25 and TF-IDF cosine. Regarding context-sensitive word embedding methods, we evaluated a multilingual version of BERT and two Portuguese focused ones: BERTimbau and BioBERT_pt. BERTimbau, a general Brazilian Portuguese-focused model, was the one which performed best among the three, ranking second in the strategies where the candidates were represented by their questions (Question-Question and Question-Question-Answer). BioBERT_pt is a Portuguese model pretrained in clinical texts, which we thought would be an advantage of the model.

However, although pretrained on texts in a domain similar to ours, the nature of the texts seems to be different. The clinical texts used to pretrain the method are more technical and focused on healthcare professionals, whereas our corpus is more related to healthcare patients and the way they pose their questions about Diabetes in social media. We also believe that the small datasets' size in which those methods were originally pre-trained did not benefit the transformer-based approaches, as been reported in the literature (Cunha et al., 2020, 2021).

Fine-tuned Approach Except for the traditional approaches, the embedding-based ones and MUSE were not trained in our benchmark, being evaluated in a learning setting called *zero-shot*. Another popular learning strategy aims to fine-tune the weights of a pre-trained large neural network, such as BERT, in a downstream task. In order to know how a fine-tuned approach would perform as a Portuguese RQE method about Diabetes, we have developed and tuned the weights of an RQE classifier based on BERTimbau. Together with MUSE and `L2R Semantic`, this fine-tuned approach (`Fine-tuned`) ranked first in the Question-Question strategy, outperforming its non-tuned version (e.g., `BERTimbau`). However, its performance lowered for the other two strategies with results similar to `BERTimbau`. We believe that the results of our fine-tuned approach was not better in these two strategies, which take candidate answers into account, due to the fact that we trim the input texts with a maximum length of 128 tokens, possibly affecting the representation of the answers in exchange of a faster performance. We leave a deeper analysis of this issue for future work.

Learn-to-rank Ensemble We also sought to investigate whether ensembling (stacking) the semantic similarity measures computed by the proposed RQE approaches could leverage better ranking results. This did not seem to be the case and, in the best situations, the ensemble had comparable results to single approaches such as BM25 and MUSE. Relying on Dalip et al. (2012), we also investigated whether quality assessment features could positively influence the ranking process. In fact, results **do not** confirm this hypothesis with `L2R All`, with semantic and quality features underperforming when compared to `L2R Semantic`, with semantic features only, in the Question-Question and Question-Question+Answer strategies.

	Q_1	Q_2	Q_3
questions	10/8	15.5/12	29/19
answers	59.5/56	86.5/73	102/96

Table 2: Length of questions and answers distributed by quartiles. Each cell contains the pair (length of failure cases/length of success cases).

Comparing Strategies Regarding the three compared strategies, results are inconclusive about which one is the best due to a high variation among the approaches. Due to the variability in terms of ranking precision between the three strategies, a choice of strategy could be made based on efficiency. In this case, the Question-Question strategy will be chosen, since, in our benchmark corpus, candidate questions have an average of 22 tokens, being much faster to process than candidate answers, with an average length of 94 tokens.

5.1 Failure Cases Analysis

We also conducted a failure cases analysis when leveraging the best-evaluated method (MUSE). A failure case happens when a non-relevant candidate question or answer is ranked at the top. Once identified, these cases were contrasted with the success cases (when a relevant candidate appears in the first position of the ranking) according to the following criteria: input length distribution, number of relevant candidates and separability.

Input Length Distribution defines the amount of processing required to understand the sentence semantics. Longer sentences require understanding more context and demand more memory to connect different concepts distributed over the sentence. In all strategies mentioned in Section 4, the length of questions and answers of the failure cases are larger than the success cases, as showed in Table 2. In the third quartile, for instance, the questions are 34.48% longer for the failure cases when compared to the success cases.

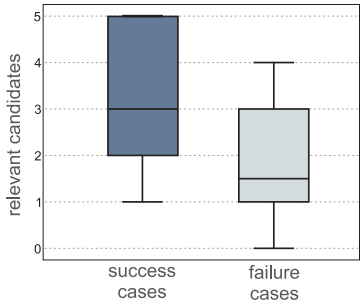

Figure 1: Number of relevant candidates per failure/success cases.

Number of Relevant Candidates the number of relevant candidates also has an important impact on the model's effectiveness since with less relevant candidates, it is more challenging to place a relevant result in the first ranking position. As demonstrated in Figure 1, the majority of failure cases (75%) had at most 3 relevant candidates with 50% of the samples having at most 1 relevant candidate. On the other hand, in success cases, there were at least 3 relevant candidates for 50% of the samples, with 25% having 5 relevant candidates.

Separability separability has to do with the ability to generate representations for similar sentences that are closer in the embedding space than semantically dissimilar sentences. We measured the amount of dispersion of similarity scores across all three strategies and the failure cases are up to 25% less separable than success cases. Since with a lower separability is harder to distinguish relevant from non-relevant candidates, the ranking produced by the model diverges from the optimal form.

Summarizing, failure cases seem to be those in which the semantic meaning is more distributed on longer sentences, containing a smaller number of relevant candidates and that are less separable.

6 Related Work

Recognizing Question Entailment (RQE) has been extensively investigated in the field so that it was included as a shared task in SemEval-2016/2017 (Task 3 - Subtask B) (Nakov et al., 2016, 2017). Under the domain of the Qatar Living corpus, the task consisted of reranking 10 candidate question-answer pairs retrieved by Google for a premise question about Qatar. Among the promising participant approaches, we highlight SimBOW (Charlet and Damnati, 2017), the winner of the shared-task in SemEval 2017. The approach works by computing the semantic similarity over the vector representations of a premise question and a corresponding candidate question-answer pair using the SoftCosine metric. Another promising participant was KeLP (Filice et al., 2016), an approach based on Tree Kernels and SVMs which provided top results in the task. After the shared-tasks, Kunneman et al. (2019) conducted a study with these approaches in order to understand the effects of particular design choices, such as the adopted preprocessing methods and word-similarity metrics.

In the medical domain, Wang et al. (2016) proposed an answer recommendation algorithm for medical community question answering. Given a user query, the system starts by looking for similar archived questions using a paragraph vector based language model (PVLM) as a similarity metric. This metric measures the distance between a premise question and a candidate one by multiplying the cosine distances among the word embedding of each word of the premise question with the paragraph vector of the archived question. In the same year, Abacha and Demner-Fushman (2016) proposed a supervised machine learning approach which classifies whether or not a candidate question can be inferred from a premise question. The questions were represented based on lexical and semantic features. More recently, Ben Abacha and Demner-Fushman (2019) proposed a siamese neural network to predict whether a candidate question is a perfect match, a relevant one or an irrelevant one to a premise question.

7 Conclusion

This study investigates Recognizing Question Entailment approaches to build a Community Question Answering about Diabetes Mellitus. Unlike previous studies in the field, ours focuses on a language other than English. Specifically, we focused on the Portuguese language. Due to the lack of resources for the language in this domain, we built a benchmark corpus, which was used to test several RQE models ranging from traditional information retrieval methods to novel large pre-trained language models and ensemble techniques using learn-to-rank techniques. Results show the power of multilingual and multi-task large neural networks such a MUSE. This sentence encoder obtained the best results of our evaluation in a zero-shot learning setup, i.e. this means it was not optimized to the target task. Results of the evaluation also show that BM25, a traditional and light information retrieval method, can obtain competitive results in the task.

Different from what was expected, state-of-art fine-tuned methods such as our BERTimbau classifier did not perform better than our MUSE zero-shot approach. We believe this may be caused by lack of training data, since our benchmark is relatively small. In future work, we plan to overcome this problem by collecting more data and expanding the corpus to other conditions related to Diabetes, such as hypertension. Like Yang et al. (2020), we also plan to augment our Portuguese training corpus by translating English questions from corpora such as MedQuad into Portuguese.

Acknowledgments

This research was partially funded by the Brazilian agencies CNPq, CAPES, and FAPEMIG. In particular, the researchers were supported by CNPQ grant No. 310630/2017-7, CAPES Post doctoral grant No. 88887.508597/2020-00, and FAPEMIG grant APQ-01.461-14.

References

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*.

Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing question entailment for medical question answering. In *AMIA Annual Symposium Proceedings*, volume 2016, page 310. American Medical Informatics Association.

Amin Ahmad, Noah Constant, Yinfei Yang, and Daniel Cer. 2019. ReQA: An evaluation for end-to-end answer retrieval models. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 137–146, Hong Kong, China. Association for Computational Linguistics.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

Delphine Charlet and Géraldine Damnati. 2017. Simbow : une mesure de similarité sémantique entre textes (simbow : a semantic similarity metric between texts). In *Actes des 24ème Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2 - Articles courts*, pages 126–133, Orléans, France. ATALA.

Washington Cunha, Sérgio D. Canuto, Felipe Viegas, Thiago Salles, Christian Gomes, Vítor Mangaravite, Elaine Resende, Thierson Rosa, Marcos André Gonçalves, and Leonardo C. da Rocha. 2020. Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. *Inf. Process. Manag.*, 57(4):102263.

Washington Cunha, Vítor Mangaravite, Christian Gomes, Sérgio D. Canuto, Elaine Resende, Cecilia Nascimento, Felipe Viegas, Celso França, Wellington Santos Martins, Jussara M. Almeida, Thierson Rosa, Leonardo C. da Rocha, and Marcos André Gonçalves. 2021. On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *Inf. Process. Manag.*, 58(3):102481.

Daniel H. Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. 2012. On multiview-based meta-learning for automatic quality assessment of wiki articles. In *Theory and Practice of Digital Libraries*, pages 234–246, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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. Association for Computational Linguistics.

Simone Filice, Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2016. KeLP at SemEval-2016 task 3: Learning semantic relations between questions and answers. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1116–1123, San Diego, California. Association for Computational Linguistics.

Nathan S. Hartmann, Erick R. Fonseca, Christopher D. Shulby, Marcos V. Treviso, Jéssica S. Rodrigues, and Sandra M. Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 122–131, Porto Alegre, RS, Brasil. SBC.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

K. Sparck Jones, S. Walker, and S. E. Robertson. 2000. A probabilistic model of information retrieval: Development and comparative experiments part 2. *Inf. Process. Manage.*, 36(6):809–840.

Florian Kunneman, Thiago Castro Ferreira, Emiel Krahmer, and Antal van den Bosch. 2019. Question similarity in community question answering: A systematic exploration of preprocessing methods and models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 593–601, Varna, Bulgaria. INCOMA Ltd.

Donald Metzler and W Bruce Croft. 2007. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 27–48, Vancouver, Canada. Association for Computational Linguistics.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 525–545, San Diego, California. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy 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. Association for Computational Linguistics.

Tony Russell-Rose and Jon Chamberlain. 2017. Expert search strategies: the information retrieval practices of healthcare information professionals. *JMIR medical informatics*, 5(4):e33.

Gerard Salton and Michael J McGill. 1986. Introduction to modern information retrieval.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jing Wang, Chuntao Man, Yifei Zhao, and Feiyue Wang. 2016. An answer recommendation algorithm for medical community question answering systems. In *2016 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, pages 139–144. IEEE.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.