Analysing Off-The-Shelf Options for Question Answering with Portuguese FAQs

Hugo Gonçalo Oliveira
CISUC, DEI, University of Coimbra, Portugal

Sara Inácio
CISUC, DEI, University of Coimbra, Portugal

Catarina Silva
CISUC, DEI, University of Coimbra, Portugal

Abstract

Following the current interest in developing automatic question answering systems, we analyse alternative approaches for finding suitable answers from a list of Frequently Asked Questions (FAQs), in Portuguese. These rely on different technologies, some more established and others more recent, and are all easily adaptable to new lists of FAQs, on new domains. We analyse the effort required for their configuration, the accuracy of their answers, and the time they take to get such answers. We conclude that traditional Information Retrieval (IR) can be a solution for smaller lists of FAQs, but approaches based on deep neural networks for sentence encoding are at least as reliable and less dependent on the number and complexity of the FAQs. We also contribute with a small dataset of Portuguese FAQs on the domain of telecommunications, which was used in our experiments.

2012 ACM Subject Classification Computing methodologies → Natural language processing

Keywords and phrases Natural Language Processing, Portuguese, Question Answering, FAQs, Information Retrieval, Sentence Encoding, Transformers

Digital Object Identifier 10.4230/OASIcs.SLATE.2022.19

Supplementary Material *Software (Source Code)*: `https://github.com/NLP-CISUC/PT_QA_Agents`

Funding This work was funded by the project POWER (grant number POCI-01-0247-FEDER-070365), co-financed by the European Regional Development Fund (FEDER), through Portugal 2020 (PT2020), and by the Competitiveness and Internationalization Operational Programme (COMPETE 2020); and by national funds through FCT, within the scope of the project CISUC (UID/CEC/00326/2020) and by European Social Fund, through the Regional Operational Program Centro 2020.

1 Introduction

As a consequence to recent advances in Artificial Intelligence and, specifically, Natural Language Processing (NLP), there has been more and more interest in the development of systems that one may interact using natural language. These include artificial agents that emulate human-to-human conversations, some capable of providing customer-support and answering complex questions (QA).

When it comes to developing such systems, a popular option is to rely on so-called Natural Language Understanding (NLU) platforms, like Google Dialogflow[1] and Microsoft LUIS[2], which offer intuitive interfaces for designing dialogue flows, and managing intents, triggered actions and answers to give. However, the configuration and the maintenance of a complete working agent often comes with great manual effort, which tends to escalate with the size and diversity of the target domain, and has to be repeated for every new agent. Moreover,

[1] `https://cloud.google.com/dialogflow`
[2] `https://www.luis.ai/`

those platforms are generally limited to one or a small set of approaches for matching user request and retrieving answers. In most cases, it will be difficult, if possible, to integrate and test alternative approaches.

A cheaper option is to develop agents that get their answers from lists of Frequently Asked Questions (FAQs), especially when such lists are available *a priori*, in websites or other documentation. The problem is then the selection of a suitable approach for matching user requests with the available FAQs. If such an approach is not dependent on the style, on the domain and on the number of FAQs, creating a new agent becomes a mater of replacing the list of FAQs with a new one.

In order to gather more information on available options for developing a system of the previous kind for Portuguese, we explore and analyse different alternative approaches. We tried to cover a range of approaches, available off-the-shelf, and adaptable to any domain. Some are more traditional (e.g., Information Retrieval, IR), and others are based on more recent transformer neural networks, for sentence encoding (USE [23], BERT [4]), then used for computing semantic similarity, for extractive QA (BERT), and for text generation (GPT2 [16]).

To analyse how the selected approaches adapt to different domains, they were tested in two datasets of FAQs, i.e., question and answer pairs, in Portuguese. Following this experiment, we provide details on the configuration effort required for each approach, on the time required by each to get answers, and on their accuracy. The latter can be computed when question variations, available for both datasets, are used for simulating user requests. One of the datasets, AIA-BDE [7], has questions on a set of subdomains of public administration and was already available for this kind of experiments. The other is on the domain of telecommunications, was created in the scope of this work, and was made available for any interested researcher.

The main conclusions were that a traditional IR approach may be enough for a smaller list of FAQs. However, transformer models for sentence encoding adapt better to larger and more complex lists. This is especially true to the model based on the Universal Sentence Encoder (USE), developed specifically for answering FAQs. Based on all the analysed aspects, results also suggest that a transformer fine-tuned for QA does not suit our goal; and that generation is also not a good option, at least if it relies on fine-tuning the original GPT2, pre-trained mostly on English text. We hope that these insights can guide those willing to build a Portuguese QA agent in Portuguese on a specific domain.

In the remainder of this paper, we overview some related work on FAQ-oriented QA. We then describe the experimentation setup, covering the approaches and the data used. Before concluding, we present and discuss the results of our analysis.

2 Related Work

Automatic Question Answering (QA) is a NLP task with the goal of obtaining answers, often out of collections of documents [22], for questions posed in natural language. Traditional approaches were based on Information Retrieval (IR) [10], but more recent approaches rely on fine-tuning transformer neural networks. For instance, when fine-tuned on datasets like SQuAD [17], those models can extract answers from limited contexts [4]. When the answer can be spread across several documents, traditional IR, namely, the BM25 ranking method, can be used for reducing the search space [12].

The problem of getting answers from FAQs is not exactly the same, because the original questions are already formulated, together with their answer. Nevertheless, IR still makes sense in this context, namely for the computing the questions, out of those available, that are most similar to user requests. Once these are identified, their answer can be given.

Work on QA based on FAQs is not recent [2] and has been attempted for different languages [9, 15, 3]. For computing the similarity between user requests and available FAQs, traditional FAQ answering systems have exploited features such as word overlap [9], synonyms [11, 15], or distributional semantic features [9, 6].

Recently, transformers were also adopted for this task. BERT has been fine-tuned for computing the request-question [13] and the request-answer [19, 13] relevance. BM25 has been used as a baseline, but may also be useful for reducing the search space, by making an initial selection of potentially-related question-answer pairs.

For Portuguese, related work resulted in the development of Amaia [20], a FAQ-answering agent that relies on a Semantic Textual Similarity model (STS) for Portuguese, trained in the ASSIN collections [5, 18], which explores a broad range of lexical, syntactic and semantic features. Amaia gets its answers from the AIA-BDE corpus [7], where several additional IR approaches were tested, including BM25 (with the Whoosh library), and others based on static word embeddings or encoding with a pre-trained BERT, multilingual and Portuguese. Reported results show that the best model depends on the kind of variation, with interesting results for BM25, word2vec embeddings, and the Portuguese BERT. More information on AIA-BDE can be found in section 3.3.

More recently, transformer models for English and Portuguese (BERT, RoBERTa, Distillbert) were fine-tuned in the ASSIN 2 collection for encoding full sentences in Portuguese and enabling to answer FAQs [3]. They were tested in a dataset of 72 FAQs, on employee assistance in a telecommunications company, some paraphrased for testing purposes. To our knowledge, this dataset is not provided. The best performance (95% accuracy) was achieved by the models that had been pre-trained for Portuguese.

This work differs from the previous because: instead of single one, it experiments with two public datasets of FAQs in Portuguese and in two different domains; it includes some approaches that had not been tested before in this scenario; and it focuses on off-the-shelf approaches, which require minimal configuration, i.e., no additional training or complex fine-tuning.

3 Experimentation Setup

This work explores a set of approaches for Automatic Question Answering (QA) based in Portuguese FAQs, in any given domain, reflected in a list of FAQs provided. This section starts by introducing the models underlying the selected approaches, then detailing these approaches and, finally, describing the data used for assessing them.

3.1 Explored Models

Explored approaches are based on different technologies, which we present here, before explaining how they were applied for QA from FAQs.

Whoosh[3] is a Python library for traditional IR which, by default, implements the BM25F probabilistic ranking function. It can be used for indexing collections of documents, according to a schema that sets available fields and their type, as well as required analysis (e.g., lowercase conversion, stemming). Given an index and a search query, Whoosh will retrieve relevant documents indexed.

[3] `https://whoosh.readthedocs.io/`

The Universal Sentence Encoder (USE) [23] is a model for encoding text in high-dimension embeddings. It is currently based on a transformer and covers 16 different languages, including Portuguese[4]. USE-QA uses these embeddings for the task of QA. It encodes pairs of sentences and context and stores them in an index built with the `simpleneighbors`[5] library. After this, given a request, USE-QA encodes it and queries the index, which returns an ordered list of approximate nearest neighbors in the semantic space.

BERT [4] is one of the most popular models based on transformers, which can be used for obtaining contextualised word or sentence embeddings, and can be further fine-tuned for different tasks like QA or natural language inference (NLI). BERTimbau [21] is a BERT model pre-trained for Portuguese which can be used for embedding words and longer sequences. This model has been fine-tuned for different tasks, including: NLI, on the ASSIN collections (hereafter, BERT-NLI); and extractive QA, in a Portuguese translation of the SQuAD 1.1 corpus (BERT-QA). The pre-trained version and the fine-tuned models are available off-the-shelf from the HuggingFace hub[6], respectively as: *neuralmind/bert-large-portuguese-cased*, *pierreguillou/bert-base-cased-squad-v1.1-portuguese*, *ricardo-filho/bert-portuguese-cased-nli-assin-assin-2*.

GPT2 [16] is another model based on the transformer architecture. However, in opposition to the previous, which use encoder blocks, GPT2 uses only decoder blocks, and can be used more like a traditional language model, i.e., it generates text from given prompts. It can be fine-tuned with text of different styles and on different domains. Using the right prompts, it has been shown to perform several tasks unsupervisedly, which is also why we decided to explore it for our purpose. Its evolution, GPT3 [1] has ten times more parameters and was pre-trained in more text, in more languages. It is known for performing well in few-shot learning. However, access to GPT3 is controlled by an API[7] and its free utilisation is limited. So, it was not used in this work.

3.2 Approaches for QA from FAQs

Different approaches were tested for QA based on FAQs, having in mind that they could be applied to any domain, regardless the availability of training data, and with the lowest possible effort. At the same time, we tried to cover approaches relying on different techniques, namely, those described in Section 3.1.

All approaches instantiate a "pipeline" that starts with a list of question-answer pairs (FAQs) and goes through two main stages: adaptation, where approach-specific preparations are performed (e.g., indexation or fine-tuning) once; and execution, where the approach waits for user queries and, upon receiving one, tries to obtain an answer, either by retrieving the most similar question or by generating the following text. We now describe how the selected models instantiate these stages.

The adaptation of Whoosh involves the indexation of the question-answer pairs. Each indexed document will have a field for the question and another for the answer. During execution, a question can be made to the index, and Whoosh will use its terms for retrieving the most similar question(s) and respective answer(s).

[4] `https://tfhub.dev/google/universal-sentence-encoder-multilingual-qa/3`
[5] `https://pypi.org/project/simpleneighbors/`
[6] `https://huggingface.co`
[7] `https://openai.com/api/`

As it happens with Whoosh, in the adaptation stage of USE-QA, question-answer pairs are indexed: questions are used as the sentences, and the full question-answer pair is used as the context. During execution, given a user query, USE-QA retrieves the most similar question indexed, together with its answer.

BERT models can be are used with the help of the HuggingFace `transformers` library[8], specifically with pipeline objects[9]. For instance, the *feature-extraction* pipeline extracts the hidden states of transformer for a given text, which can be used as its embeddings. It is used with the pre-trained version of BERTimbau (large), in order to get the `[CLS]` vector. From here, we refer to this approach as BERT-FE.

BERT-NLI, on the other hand, is a sentence transformer, and thus more suitable for encoding full sequences of text, besides having been trained for a task where representing the meaning of sentences is important (i.e., NLI). Although it could be used with the simple `transformers` library, the `sentence-transformers`[10] library is more suitable for this model. It is loaded via the SentenceTransformer object, where embeddings can be obtained through the *encode* method.

In the adaptation stage of both BERT-FE and BERT-NLI, the models are loaded and the questions in the list of FAQs are encoded. In the execution stage, given user queries are encoded and the most similar questions, i.e., those maximising the cosine similarity, are computed and retrieved.

BERT-QA is also used with a pipeline object, this time of the *question-answering* type. This model expects a textual context where it will search for answers to given questions. Ideally, for our use case, the full list of FAQs would be used. However, there are limitations on the size of the context (i.e., maximum 512 tokens for BERT-base). So, to use it, the list of FAQs has to be first split into smaller subsets. To do this, and to keep the approach independent of the list of FAQs, its adaptation stage includes the additional step of clustering the FAQs with k-means, hoping to discover groups of related FAQs, when represented by BERT-FE embeddings. To select the number of clusters, we relied on the Elbow [8] method, which returned the optimal number in the 1–100 interval.

For the adaptation of GPT2, the model is fine-tuned with the lists of original questions and their answers. This was done with the `gpt-2-simple` library[11]. We used GPT2 medium, which has 355M parameters, and set a temperature of 0.2, for avoiding highly random text. For a model like GPT2, fine-tuning does not require the definition of a task. The only requirement is to have data on the style of the text to generate, in this case, the list of FAQs and their answers. Expectations were that, when prompted with a question, GPT2 would generate a suitable answer from what it "had seen" during fine-tuning. Given the format of the lists of FAQs (see Section 3.3), for generation, we used a prefix that started with `P:`, followed by the user query and by `R:`. In addition, we define '`\n\n`' as the truncation token because, before fine-tuning, we force that every question-answer is followed by an empty line.

3.3 Data

Since we also wanted to confirm that the selected approaches would work in different domains, we tested them in two different datasets with Portuguese FAQs and their variations, of different sizes and on different domains.

[8] `https://huggingface.co/docs/transformers`
[9] `https://huggingface.co/docs/transformers/main_classes/pipelines`
[10] `https://huggingface.co/sentence-transformers`
[11] `https://github.com/minimaxir/gpt-2-simple`

The AIA-BDE corpus [7] contains 855 question-answer pairs on topics related to public administration. For each pair, variations were produced by different approaches and groups of people. Variations simulate user requests by paraphrasing the original questions, using different words, sometimes with missing information. QA approaches can be tested in the task of retrieving the original question, and the associated answer, given its variations. In AIA-BDE, questions are in lines starting with `P:`, their answers are in the following line that starts with `R:`, and variations are between them, in lines starting with `VX:`, where `X` depends on the kind of variation. See Figure 1 for an example of a question in AIA-BDE, followed by manually-created variations and the answer.

```
P:Como pedir o Cartão Provisório de Identificação de Pessoa Coletiva?
VUC:Como posso obter o cartão provisório de identificação de pessoa coletiva?
VUC:Onde posso pedir o cartão provisório de pessoa coletiva?
VIN:Como posso pedir o Cartão provisório de identificação de pessoa coletiva?
VIN:Qual o procedimento para obter o Cartão Provisório de Identificação de
Pessoa Coletiva?
R:O Cartão Provisório de Identificação de Pessoa Coletiva deixou de ser emitido, (...)
Atualmente, existe apenas o Cartão da Empresa e o Cartão de Pessoa Coletiva, que são
emitidos para entidades definitivamente registadas ou inscritas.
```

Figure 1 Example of a FAQ in AIA-BDE, its variations and answer.

A second dataset (hereafter, Telecom) was produced specifically for this work. It is on a different domain, but its creation followed a similar approach to the creation of AIA-BDE. Telecom has 172 question-answer pairs gathered from various online sources, covering instruction manuals of telecommunication equipment and services by the Portuguese telecommunications operator MEO[12]. Besides being on a different domain, the style of Telecom FAQs is significantly different than AIA-BDE. For instance, it includes several questions that are, in fact, stating issues (see Figure 2 for examples). For its creation, and similarly to AIA-BDE, at least two question variations were manually created, by two people, for 103 of the FAQs in the dataset. In this case, they were included in a different file (see Figure 3 for the variations of the questions in Figure 2).

```
P: Internet com quebras, com ligação por cabo
R: Desligue todos os equipamentos da linha telefónica ou da rede sem fios.  Depois, ...

P: Sistema Operativo Mac OSX: Como ligar a uma rede sem fios
R: Procure as redes sem fios Clique no ícone de WiFi e procure as redes sem fios ...

P: Como definir um código pessoal de acesso ao Voice Mail?
R: Se aceder ao Voice Mail a partir do seu próprio telefone, não é necessário colocar o
código pessoal.
```

Figure 2 Example of FAQs in the Telecom dataset.

4 Analysis

The selected approaches were configured for answering questions based on the FAQs in the Telecom and AIA-BDE datasets. This section analyses the configuration effort involved, the accuracy of the given answers, and the time taken for answering. The former and the latter are analysed subjectively, while the available question variations can be used for quantifying accuracy.

[12] `https://www.meo.pt/ajuda-e-suporte/produtos-meo/internet/equipamentos`

```
Ligação por cabo de internet com quebras
Internet com falhas com ligação por cabo

Como fazer ligação a uma rede sem fios num Mac OSX?
Como ligar a uma rede sem fios com o sistema operativo mac OSX?

Como posso definir um código de acesso ao voice mail?
Como defino um código de acesso ao Voice Mail?
```

Figure 3 Example of variations for the Telecom dataset.

4.1 Configuration Effort

Selected approaches are quite different and thus require a different configuration effort. This is probably the most subjective aspect analysed but, for those that are not familiar with any of the required technologies, it can be as important as the other two. Our analysis is based on a brief description of the steps required for having each approach ready to work, i.e., answering questions.

Thanks to the Huggingface `transformers` and `sentence-transformers` libraries, transformers are nowadays straightforward to use. In just a few lines, we can start using BERT and computing embeddings. After this, it is just a matter of computing cosines. Therefore, we would say that the BERT-FE and the BERT-NLI approaches are those that require less configuration effort. If it were not for the necessary clustering, BERT-QA would require a similar effort. However, this step increases the complexity of the configuration.

The effort of using USE-QA is also low. Some time was required for better understanding the input format for the list of FAQs. However, every step detailed in documentation[13].

GPT2 could also be used with the `transformers` library, but the tested approach requires fine-tuning, which is not as straightforward as using an available model. However, in this case we used the `gpt-2-simple` library, which is also well-documented and makes it straightforward to fine-tune the original GPT2 models and use the result.

Using Whoosh requires more steps than the previous models, namely for defining the index schema, creating the index, searching and querying, but every step is also documented[14]. In the end, we can say that the required effort is comparable to using GPT2.

4.2 Answering Accuracy

Since both datasets include question variations, which can simulate user requests, each approach was assessed by the answers given for those variations. For this experiment, Tables 1 and 2 report on different measures. Accuracy is the proportion of variations for which the answer was the same as the one for the original question. This is computed for only the variations (Acc-vars)[15] and also when the original questions are included (Acc-all). However, simply comparing the answers is unfair for GPT2, because it may generate answers that are not exactly the same as those expected, even if they might transmit a close meaning. Therefore, we include two additional metrics: BLEU [14], which measures the surface text similarity, and BERTScore [25], which, in an attempt to consider the meaning of text, relies on contextual embeddings, in this case, obtained from BERTimbau.

Figures show that overall accuracies are lower in the AIA-BDE corpus. This was expected because AIA-BDE is larger, meaning that there are many more possibilities when searching for similar questions. Another contribution to this may result from the variations of AIA-BDE,

[13] `https://www.tensorflow.org/hub/tutorials/retrieval_with_tf_hub_universal_encoder_qa`
[14] `https://whoosh.readthedocs.io/en/latest/quickstart.html`
[15] For AIA-BDE, only the manually-produced variations were considered.

Table 1 Performance in Telecom FAQs.

Approach	Acc-all	Acc-vars	BLEU	BERTScore
Whoosh	91.53%	88.35%	94.76	98.63
USE-QA	91.53%	88.29%	95.30	99.00
BERT-FE	66.40%	38.83%	79.58	97.38
BERT-NLI	74.60%	54.15%	84.16	97.58
BERT-QA	0.00%	0.00%	0.12	88.71
GPT2	0.00%	0.00%	22.11	90.12

Table 2 Performance in AIA-BDE FAQs.

Approach	Acc-all	Acc-vars	BLEU	BERTScore
Whoosh	70.22%	65.59%	70.45	94.34
USE-QA	84.37%	79.35%	88.63	97.85
BERT-FE	75.05%	68.42%	81.04	96.48
BERT-NLI	79.73%	71.35%	84.44	97.09
BERT-QA	0.00%	0.00%	1.00	74.19

which, at the surface level, can be more different from the original questions. Yet, regardless of the dataset, top scores are always achieved by USE-QA, which seems to be the best option for our purpose.

In the Telecom dataset, the performance of USE-QA is matched by the best configuration of Whoosh. We only show the performance of this configuration, which is based on defaults plus a Portuguese Analyzer. Other configurations were tested (e.g., 3 and 4-grams), but differences were minimal. The performance of Whoosh suggests that, in some cases, traditional IR can be enough. However, we also see that its performance drops for AIA-BDE, a larger dataset and possibly more difficult.

BERT-NLI, fine-tuned for representing the meaning of sentences, performs better than the pre-trained BERT, which was somehow expected. In opposition to the other approaches, the performance of BERT-NLI increases in AIA-BDE, where it has the second best accuracy, confirming that the model representations go beyond surface text, and thus handle well variations using different words.

BERT-QA always extracts spans of text that have nothing to do with the question, which results in 0 accuracy and leads to the conclusion that it is not an option for this task. This is a consequence of both: (i) noise when selecting the cluster of FAQs to use as context; and, especially, (b) the format of the datasets (question-answer), which ends up being different from SQuAD (context-question-answer). While, in the future, we could experiment with a different clustering algorithm, a different numbers of clusters, and a different representation of FAQs, there is not much we can do about the latter, other than fine-tuning BERT on a dataset with a closer style.

We also confirm that GPT2 cannot generate answers that completely match the original ones[16]. Yet, even considering BLEU, it is way below the other approaches. GPT2 was fine-tuned in Portuguese data, but its starting point was the original GPT2 (medium), pre-trained mostly on English text, which might have had an impact here. In the future, we should try fine-tuning a GPT2 pre-trained for Portuguese[17], or move on to recent open alternatives of GPT3 (e.g., Meta OPT [24]).

[16] Due to lack of memory, it was not possible to compute the figures for GPT2 in AIA-BDE
[17] `https://huggingface.co/pierreguillou/gpt2-small-portuguese`

Apart from enabling a comparison with GPT2, BLEU and BERTScore do not add much to our conclusions. BERTScore is always high, even for completely different texts, and should only be considered relatively. It might also be positively biased towards BERT-FE, which can be a consequence of using the same model for computing the scores.

4.3 Time

We analyse the time required for the adaptation and the execution stages. Our experiments were run in Google Colab[18] without hardware acceleration, and average times were measured with the `time` library.

Adaptation, which only has to be done once for each run, includes embedding (BERT, USE-QA) and indexing (Whoosh) the questions, clustering (BERT-QA), or fine-tuning (GPT2). The latter is by far what takes more time in this stage, i.e., more than 1 hour for fine-tuning GPT2 with the Telecom dataset. For all the others, adaptation takes between 1 second (Whoosh) and 2 minutes (BERT-QA) in Telecom, and between 5 seconds (Whoosh) and 9 minutes (BERT-QA) in AIA-BDE.

Regarding the execution stage, we measure the average time taken between giving a question as input and getting its answer. Here, most approaches take less than 1 second in both datasets, with the exceptions being again GPT2 (40 seconds) and BERT-QA (30 to 90 seconds, depending on the number of clusters and size of the selected).

5 Conclusion

We have tested several approaches for automatic QA based in Portuguese FAQs, relying on different techniques, but all easily adapted to any domain. Following the results obtained in two datasets of FAQs and their variations, we can immediately discard two approaches, due to issues on accuracy and configuration effort, namely BERT-QA and GPT2. As for the others, USE-QA, based on the Universal Sentence Encoder, revealed to be the best option for answering FAQs. We should nevertheless highlight the performance of the BM25F traditional IR method, which performs well, especially in the smaller dataset. BERT-NLI is not a bad option either, but BERT would probably benefit from fine-tuning on data in the target style or domain. This was not considered for this work because we wanted to rely, as much as possible, on what was available off-the-shelf. However, it is something to try in the future, as well as further experiments with GPT2, GPT3 and comparable models.

Following the current interest in developing conversational agents and QA systems, often in domains for which FAQs are already available, the reported experiments may help those planning to develop or upgrade such a system. In the meantime, the code used for our experimentation, as well as the new Telecom dataset, are publicly available from a Github repository[19].

References

1 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[18] `https://colab.research.google.com/`
[19] `https://github.com/NLP-CISUC/PT_QA_Agents`.

2 Robin D Burke, Kristian J Hammond, Vladimir Kulyukin, Steven L Lytinen, Noriko Tomuro, and Scott Schoenberg. Question answering from frequently asked question files: Experiences with the FAQ finder system. *AI magazine*, 18(2):57–57, 1997.

3 Nuno Carriço and Paulo Quaresma. Sentence embeddings and sentence similarity for portuguese faqs. *Proceedings of IberSPEECH 2021*, pages 200–204, 2021.

4 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

5 Erick Fonseca, Leandro Santos, Marcelo Criscuolo, and Sandra Aluísio. Visão geral da avaliação de similaridade semântica e inferência textual. *Linguamática*, 8(2):3–13, 2016.

6 Erick R. Fonseca, Simone Magnolini, Anna Feltracco, Mohammed R. H. Qwaider, and Bernardo Magnini. Tweaking word embeddings for faq ranking. In *Proceedings of 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 1749. CEUR-WS, 2016.

7 Hugo Gonçalo Oliveira and Ana Alves. AIA-BDE: um corpo de perguntas, variações e outras anotações. *Linguamática*, 13(2):19–35, December 2021.

8 Kalpana D Joshi and PS Nalwade. Modified k-means for better initial cluster centres. *International Journal of Computer Science and Mobile Computing*, 2(7):219–223, 2013.

9 Mladen Karan, Lovro Žmak, and Jan Šnajder. Frequently asked questions retrieval for Croatian based on semantic textual similarity. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 24–33, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

10 Oleksandr Kolomiyets and Marie-Francine Moens. A Survey on Question Answering Technology from an Information Retrieval Perspective. *Information Sciences*, 181(24):5412–5434, December 2011.

11 Govind Kothari, Sumit Negi, Tanveer A. Faruquie, Venkatesan T. Chakaravarthy, and L. Venkata Subramaniam. Sms based interface for faq retrieval. In *Proc Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, ACL '09, pages 852–860. ACL, 2009.

12 Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy, July 2019. ACL.

13 Yosi Mass, Boaz Carmeli, Haggai Roitman, and David Konopnicki. Unsupervised FAQ retrieval with question generation and BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 807–812, 2020.

14 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

15 Arianna Pipitone, Giuseppe Tirone, and Roberto Pirrone. ChiLab4It system in the QA4FAQ competition. In *Proceedings of 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 1749. CEUR-WS, 2016. URL: `http://ceur-ws.org/Vol-1749/`.

16 Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

17 Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 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, 2016.

18 Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. The ASSIN 2 shared task: a quick overview. In *Computational Processing of the Portuguese Language - 13th International Conference, PROPOR 2020, Évora, Portugal, March 2-4, 2020, Proceedings*, volume 12037 of *LNCS*, pages 406–412. Springer, 2020.

19 Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. FAQ retrieval using query-question similarity and BERT-based query-answer relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1113–1116, 2019.

20 José Santos, Luís Duarte, João Ferreira, Ana Alves, and Hugo Gonçalo Oliveira. Developing Amaia: A conversational agent for helping portuguese entrepreneurs — an extensive exploration of question-matching approaches for Portuguese. *Information*, 11(9), 2020.

21 Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, 2020.

22 Ellen M. Voorhees. The TREC-8 Question Answering track report. In *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA*. NIST, November 1999.

23 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. 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. ACL, July 2020.

24 Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint*, 2022. `arXiv:2205.01068`.

25 Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. *arXiv preprint*, 2019. `arXiv:1904.09675`.