Sentence Embeddings and Sentence Similarity for Portuguese FAQs

Nuno Carriço[1], *Paulo Quaresma*[2]

[1,2]Computer Science Department, University of Évora

nfmc@uevora.pt, pq@uevora.pt

Abstract

Virtual Assistant Bots are becoming essential in business models. This aims to provide customer service without the need of a human operator. Thus, the first step is to understand what a customer needs. To achieve this, we compute the sentence distance between a set of predefined FAQs and the user sentence, and extract the closest FAQ. While the problem has satisfactory results for English language, it is not the case for Portuguese language. Therefore, we propose the use of portuguese BERT models to obtain the sentence embeddings of both the FAQs and user sentence, in order to compute their distances scores. The BERT models are fine tuned with the ASSIN 2 dataset for sentence similarity tasks to achieve better performance. The fine tuned models were evaluated against ASSIN 2 test set.

The FAQs embeddings are inserted in a FAISS index, which is used to extract the n closest FAQs embeddings to a user sentence. The index provides an efficient way to maintain the embeddings and search for the closest neighbors given a query data point. Given the set of FAQs, we built sample user questions, labelled with their corresponding FAQ, to test the setup.

Index Terms: BERT, sentence embeddings, sentence similarity, paraphrase searching

1. Introduction

Virtual Assistant Bots are becoming an essential part of a company customer service. A customer is able to receive support at any time without the need of directly contact a company's assistance line and wait for a response. Furthermore, these assistants are now evolving in the direction of more personalized support and can be easily deployed to a variety of different tasks. Nonetheless, the first step to all of this is understanding. Our virtual assistant must be capable of understanding, to some degree, what the customer's intention is. For that matter, the collection of most common problems/questions (FAQs) a customer might encounter is a valuable resource to start with. These FAQs will serve as a base to compare user sentences as a way to "understand" what the customer needs. Therefore, we propose a simple and scalable system capable of efficiently computing the closest FAQ to a user sentence.

1.1. Related Work

Word embeddings has been a wide research topic. Various methods of obtaining word embeddings have been successful, such as ELMo [1] , which enriches word embeddings using internal network layer information, or BERT [2] with successive application of attention mechanisms to extract relations between words in a given sentence. Other methods include positional dependency to obtain the embeddings [3], that is, a triplet $(w_t,\ c,\ w_c)$ is used to construct the embeddings, where w_t, w_c, c represent, respectively, the target word, the context word and the positional dependency computed from the context.

Sentence encoders have also been in study and are applied succefuly in numerous natural language processing tasks. These encoders range from simple LSTM networks to self attention networks and hierarchical convolutional networks [4]. Recently, BERT embeddings have been in use to produce meaningful representations of sentences with the assistance of siamese networks [5].

Semantic search engines were taken a step further using word embeddings. [6] used ELMo as a base to build a word embedding based search engine. The procedure is simple and works as follows:

1. Input a query sentence and obtain it's embedding using ELMo.
2. Compute the cosine similarities scores between the previous embedding and the remaining embeddings.
3. Return the n closest vectors.

However, as the number of vectors in the search space grows, the complexity of the search also increases. Similar approaches were taken by [7] which makes BERT available as service. In this case, the embeddings are obtained with BERT, and the scores computed using the normalized dot product between the two embedding vectors. This approach also suffers from the search complexity problem. With our proposed framework, we make use of sentence embeddings to grasp the general meaning of the whole sentence. Also, the navigation trough the search space is done using an index as a way to reduce search speed.

1.2. Framework

The proposed framework consists of two sub tasks. First, we need to represent both the FAQs and user sentences by a vector which encodes their semantic values. Second, given the sentence's representation, we need to compute their similarity. This will allow us to pick the most likely FAQ that satisfies the customer intention.

Regarding the encoding step, a variety of options to build word embeddings were already available, such as Word2vec [8], GloVe [9], ELMo [1] and BERT [2]. However, we intended to encode the whole sentence, not just the relations between the words. For this reason, we decided to use Sentence BERT (SBERT) [5] as our transformer. SBERT computes the sentence embedding in the following manner: first, it computes the word embeddings using a BERT model; to the previous output, it applies a pooling layer to build the sentence embedding. For the pooling layer, we have the following strategies:

- Computing the mean of the output vectors.
- Computing the max of the output vectors.
- Using the CLS token.

Moreover, the underlying BERT model is fine tuned with siamese and triplet networks with the purpose of producing meaningful sentence embeddings that can be easily compared.

10.21437/IberSPEECH.2021-43

Since our proposed framework is expected to attent online requests from customers, we need an efficient way to search the FAQ search space for the closest FAQ to a user sentence using a given metric for similarity. For this step we use FAISS [10]. FAISS is an open source library developed by Facebook AI research group. It is mainly used for efficient space search and clustering of vectors. FAISS builds a data structure, an index, representing a given set of vectors. When a query vector is introduced in the search space, FAISS efficiently computes the distances between the query vector and the remaining index vectors, returning the k-closest index vectors. FAISS makes available multiple types of indexes. For the problem at hand, we chose the indexes IndexFlatL2 and IndexFlatIP, which use the euclidean distance and inner product, respectively, as distance metrics.

A preview of the system's architectures is displayed in figure 1.

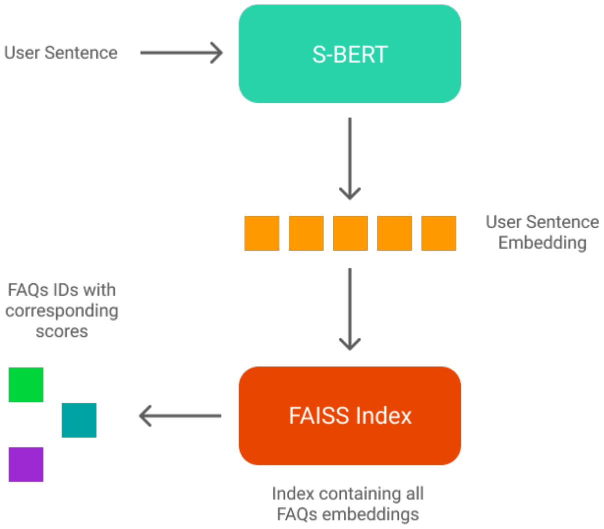
User Sentence
S-BERT
User Sentence Embedding
FAQs IDs with corresponding scores
FAISS Index
Index containing all FAQs embeddings

Figure 1: *Proposed Framework Architecture*

1.3. FAQ Corpora

The FAQ corpora was provided by a Portuguese telecommunications company and contains a total of 72 FAQs, ranging over topics related to employee assistance. These topics include matters such as, holidays, supplementary work, family, communication benefits, meals and retirement. Examples of such questions are presented in Table 1.

1.4. ASSIN 2 Dataset

ASSIN 2 [1] corpus consists of simple sentences that do not include named entities or indirect speech and all verbs are in the present tense. The training set is composed of 6500 sentence pairs, while the validation set contains 500 sentence pairs. These sentences are manually annotated for text entailment and sentence similarity. The semantic similarity values range over 1 to 5 and the text entailment classes are one of entailment or none. For the test set, it contains roughly 3000 sentence pairs, following the same annotation schema.

For the evaluation step, F1 of precision and recall is used to evaluate text entailment, and Pearson correlation is applied to the semantic similarity task. Below we provide the scores of the highest performing contestants for the semantic similarity task.

Table 1: *Example Questions in FAQ corpora*

Topic	Example Question
Communication Benefits	What are the communication benefits after retiring?
Family	Regarding family assistance, who supports the payment?
Meals	What is the processing rule of meal allowance?
Holidays	Is it allowed to anticipate holidays?
Retirement	How to request pension support?

Table 2: *ASSIN 2 Similary Task Results*

Team	Pearson	MSE
IPR	0.826	0.52
IPR	0.809	0.62
L2F/INESC	0.778	0.52
Stilingue	0.800	0.39
Stilingue	0.817	0.47

2. Experiments

2.1. FAQ Test Set

For the given FAQ corpora, we were not provided with user queries, therefore, we developed a simple set of possible user questions. For each topic in the corpora, we chose two questions and rephrased them, ending with a total of 24 questions. Examples are provided in Table 3. For the purposes of testing, each rephrased question is labelled with its correct/original FAQ.

2.2. Pre Trained

Based on the performance on the FAQ test set, we filtered out pre trained models that wouldn't be worth fine tuning. The pre trained BERT models tested were the following: Portuguese language BERT models [11], namely bert-base-portuguese-cased (BBPC) and bert-large-portuguese-cased (BLPC); English language BERT models, that is, roberta-base-nli-stsb-mean-tokens (Roberta-BNS-Mean), roberta-large-nli-stsb-mean-tokens (Roberta-LNS-Mean) and distilbert-base-nli-stsb-mean-tokens (Distilbert-BNS-Mean).

The procedure worked as follows:

1. Using SBERT, we computed the sentence embeddings of both the user sentences and the FAQs, using HugingFace Transformers [12].

[1] https://sites.google.com/view/assin2/english

Table 3: *Example Question Rephrasing*

Original FAQ	Rephrased FAQ
What are the conditions to acquire equipment in a store?	How can equipment be acquired?
Is it allowed to anticipate holidays?	Can I anticipate holidays?
How many hours of effective work are necessary in order to request meal allowance?	How many hours do I have to work to request meal allowance?

2. By applying cosine similarity to each pair of user sentence embedding and FAQ embedding, we obtained the similarity scores.
3. Finally, for each user query, we picked the FAQ with the highest similarity score as the predicted FAQ.

To evaluate the results, we compute the accuracy of the predicted FAQs. The preliminary results for the best pre trained models are as follows in Table 4

Table 4: *Preliminary Testing*

Model	Accuracy
BBPC-Mean	**0.875**
BBPC-Max	0.791
BLPC-Mean	0.75
BLPC-Max	0.833
Distilbert-BNS-Mean	0.833
Roberta-BNS-Mean	**0.875**
Roberta-LNS-Mean	0.791

From the results presented in Table 4, the models BBPC-Mean and Roberta-BNS-Mean produce the best results with an accuracy of 0.875. However, since Roberta-BNS-Mean is an English language model, we do not proceed to fine tuned it with the ASSIN 2 dataset.

2.3. Trained

Given that our goal was to build a setup for Portuguese language FAQs, we picked the best Portuguese language models from the preliminary results, namely BBPC-Mean/Max and BLPC-Mean/Max. These models were further fine tuned using ASSIN 2 dataset for 10 epochs using a batch size of 16 with cosine similarity loss as the loss function. We also experimented with different post processing layers, including CNN, LSTM and DAN.

Testing these fine tuned models against the FAQ test set, we got the results presented in Table 5.

Based on the results, we chose one model for each of the BBPC and BLPC base models, in this case, BBPC-Mean-CNN and BLPC-Mean. Models without post processing were favoured over those with post processing layers, in case of equal accuracy. These two models were further tested against ASSIN 2 test set, yielding the results in Table 6.

Table 5: *Trained Models Results*

Model	No Post Processing	CNN	LSTM	CNN-DAN
BBPC-Mean	0.875	**0.917**	0.833	0.875
BBPC-Max	**0.875**	0.875	0.833	0.875
BLPC-Mean	**0.917**	0.917	0.875	0.875
BLPC-Max	0.75	**0.875**	0.833	0.875

Table 6: *ASSIN 2 Comparison*

Model/Team	Pearson	MSE
BBPC-Mean-CNN	0.822	0.8
BLPC-Mean	**0.831**	0.78
IPR	0.826	0.52
IPR	0.809	0.62
L2F/INESC	0.778	0.52
Stilingue	0.800	**0.39**
Stilingue	0.817	0.47

Looking at Table 6, we observe that the fine tuned model BLPC-Mean outperformed the best contestant by a factor of 0.05. Nonetheless, the model BBPC-Mean-CNN also keeps on pair with the majority of the remaining contestants. Both of these could be improved to reduce the MSE scores, which are the highest among the given results.

2.4. FAISS Indexes

The setup is required to run online, that involves computing efficiently the most similar FAQ to a user query. Moreover, we kept in mind that FAQs change over time, so it could necessary to add or remove FAQs from the search space. For that matter, we introduce FAISS indexes to our framework.

First, we compute all the FAQs sentence embeddings and store them in a FAISS index. In case of needing a new FAQ, we simply coompute its sentence embedding and insert it into the index. The removal is just as easy, since we can attribute an id to each vector in the index, we can simply remove an embedding using its corresponding id.

When a new user sentence arrives, it is computed its sentence embedding and we query the index for the k closest FAQs embeddings, returning it's ids.

To test this approach, we used the developed FAQ set together with two FAISS indexes: IndexFlatL2 and IndexFlatIP. The results are shown in Table 8

Table 7: *Sample Results*

Query	BBPC-Mean-CNN	BLPC-Mean	BBPC-Mean-CNN IndexFlatIP	BLPC-Mean IndexFlatL2
Posso usufruir da dispensa semanal em cursos sem componente letiva?	Qual o código para gozo de férias por trabalho suplementar? ✗	Qual o código para gozo de férias por trabalho suplementar? ✗	Qual o código para gozo de férias por trabalho suplementar? ✗	Qual o código para gozo de férias por trabalho suplementar? ✗
Como pedir benefícios de comunicações?	Como requerer/transferir benefícios comunicações colaborador para uma nova conta/linha de rede (telefone)? ✓	Quais os benefícios de comunicações na reforma? ✗	Os trabalhadores não ativos podem usufruir do cartão Galp frota colaboradores? ✗	Como requerer/transferir benefícios comunicações colaborador para uma nova conta/linha de rede (telefone)? ✓
Que taxa de IRS aplica a empresa?	Qual o valor das ajudas de custo nacional/estrangeiro? ✗	Que taxa de IRS me foi aplicada pela empresa? ✓	Que taxa de IRS me foi aplicada pela empresa? ✓	Que taxa de IRS me foi aplicada pela empresa? ✓

Table 8: *Index Results*

Model	IndexFlatL2	IndexFlatIP
BBPC-Mean	0.875	0.833
BBPC-Mean-CNN	0.875	**0.917**
BBPC-Max	0.833	0.5
BBPC-Max-CNN	0.875	0.208
BLPC-Mean	**0.958**	**0.917**
BLPC-Mean-CNN	0.917	0.875
BLPC-Max	0.75	0.875
BLPC-Max-CNN	0.833	0.667

From Table 8, for the IndexFlatL2 index, the model BLPC-Mean outperformed the remaining models with an accuracy of 0.958. Focusing our attention on the IndexFlatIP index, the best performing models are BBPC-Mean-CNN and BLPC-Mean, both having an accuracy of 0.917. Based on these results, the model BLPC-Mean together with IndexFlatL2 produces the best setup.

2.5. Results

For the developed test set, the models, in general, select the correct FAQ in the majority of the cases, keeping the accuracy above 80%, as we already presented in tables 5 and 8. However, there are some sentences for which the models did not find the similarity. Looking at Table 7, we depicted the most problematic queries. For the first query, no model found the correct FAQ, which, in this case, was the text *Nos cursos em que não existe componente letiva (frequência de aulas), designadamente, em mestrados e doutoramentos é possível beneficiar da dispensa semanal prevista no estatudo de trabalhador - estudante ?*. The lack of fine tuning with FAQ domain data could be causing this issue, resulting in distant embeddings for this query and respective FAQ. Regarding the two other queries, the use of indexes generally helps finding the correct FAQ, specially, the model BLPC-Mean for the second query, and the model BBPC-Mean-CNN for the third query.

It would be needed a much larger dataset to assess how the system scales. Nonetheless, as of the time of writing, the system is currently being tested in such scenarios by the requesting company, and soon we will be able to provide results on that matter.

3. Conclusion

With this paper, we aimed at developing a framework that could serve as a baseline for online sentence similarity search, in this case, for virtual assistant bots and FAQs similarity. For the Portuguese language, the results show that the BLPC-Mean model together with the IndexFlatL2 index are the more promising combination for the task at hand. Nonetheless, this framework can be extended to any language simply by switching the underlying SBERT model to a model of the desired language.

Future work could involve applying dimensional reduction techniques and testing new underlying models using recent research, such as tBERT. Additionally, considering the FAQ's answers to find the correct FAQ would be of interest as well. Moreover, instead of using FAISS indexes, it could be of use trying different clustering methods to extract the closest FAQ to a sentence. Also, it would be useful to develop a full Portuguese language dataset for FAQ search as a base reference for future projects.

4. References

[1] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," 2018.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-

training of deep bidirectional transformers for language understanding," 2019.

[3] Y. Yin, C. Wang, and M. Zhang, "Pod: Positional dependency-based word embedding for aspect term extraction," 2020.

[4] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," 2018.

[5] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[6] "Elmo: Contextual language embedding," https://towardsdatascience.com/elmo-contextual-language-embedding-335de2268604, accessed: 2020-12-20.

[7] "Bert as a service," https://github.com/hanxiao/bert-as-service#building-a-qa-semantic-search-engine-in-3-minutes, accessed: 2020-12-20.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[9] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[10] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *arXiv preprint arXiv:1702.08734*, 2017.

[11] F. Souza, R. Nogueira, and R. 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.

[12] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6