# Bootstrapping a Data-Set and Model for Question-Answering in Portuguese

Nuno Ramos Carvalho
Rua A 350 2E, 4810-217 Guimararães, Portugal

Alberto Simões
2Ai, School of Technology, IPCA, Barcelos, Portugal

José João Almeida
Centro Algoritmi, Departamento de Informática, University of Minho, Braga, Portugal

## Abstract

Question answering systems are mainly concerned with fulfilling an information query written in natural language, given a collection of documents with relevant information. They are key elements in many popular application systems as personal assistants, chat-bots, or even FAQ-based online support systems.

This paper describes an exploratory work carried out to come up with a state-of-the-art model for question-answering tasks, for the Portuguese language, based on deep neural networks. We also describe the automatic construction of a data-set for training and testing the model.

The final model is not trained in any specific topic or context, and is able to handle generic documents, achieving 50% accuracy in the testing data-set. While the results are not exceptional, this work can support further development in the area, as both the data-set and model are publicly available.

**2012 ACM Subject Classification** Computing methodologies → Discourse, dialogue and pragmatics

**Keywords and phrases** Portuguese language, question answering, deep learning

**Digital Object Identifier** 10.4230/OASIcs.SLATE.2021.18

**Category** Short Paper

**Supplementary Material** *Software (Source Code)*: `https://github.com/nunorc/qaptnet`

**Funding** This project was partly funded through the European Regional Development Fund (FEDER), by Portuguese national funds (PIDDAC), through the FCT – Fundação para a Ciência e Tecnologia and FCT/MCTES under the scope of the projects "UIDB/05549/2020" and "UIDB/00319/2020"

## 1 Introduction

Modern applications use different channels of communication to exchange information with their user base, from businesses to governmental organizations. The increase of information available in digital format has resulted in a high demand for effective, continuous and uninterrupted, information provision services. In this context, tools capable of answering simple questions in real time, without human intervention, are currently in high demand. Some examples of these tools are commonly known as *chatbots* or digital assistants [4].

In the last years there has been a large amount of research in the development of these applications, mostly motivated by the increase of available data, recent advances in Machine Learning techniques, namely on Deep Neural Networks or Word Embeddings, and the increase of computational power readily available. This also spawned some frameworks providing *out-of-the-box* applications that enable a quick, and cheap, implementation of such systems with an acceptable quality.

10th Symposium on Languages, Applications and Technologies (SLATE 2021).
Editors: Ricardo Queirós, Mário Pinto, Alberto Simões, Filipe Portela, and Maria João Pereira; Article No. 18; pp. 18:1–18:5
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

A common key element in these frameworks, and corresponding workflows, is a question answering model. These models are able to find an answer to an information query, described by the user in natural language, from a collection of texts. Deep learning neural networks [5] currently achieve some of the best results for this specific task [3, 2, 9].

This paper introduces an exploratory work carried out to build a model for performing question answering tasks for the Portuguese language. Although the current literature is rich on models and techniques, some languages (in particular, the Portuguese language) lack state-of-the-art implementations of such models. Therefore, the guiding research question for this work is defined as:

> Can transfer learning be explored to train a model, for the Portuguese language, capable of providing satisfying results for finding an answer to a information query writen in natural language from a collection of texts?

Given the lack of computational power and enough data to train a model from scratch, and also in order to reuse information obtained from previously training in other languages, we start the work with the parameters from a pre-trained model. This means that we take advantage of the parameters exploration during previous training, and fine tune the model for the Portuguese language. Transfer learning encompasses the idea of not starting the model training stage from scratch, bootstrapping the model using parameters from previously training steps [7]. This usually allows the training process to achieve better results with less data or fewer training steps.

The goal of the model is: given a snippet of text and an information query written in natural language, find the answer to the query in the text. This snippet of text is usually referred to as a context, and it is assumed that the given context contains the answer to the query. This enables the implementation of *chatbots* (or similar applications) that can answer questions systematically without being explicitly programmed to do so, and independently of the subject or topic at hand.

This paper is organized as follows: Section 2 introduces the data-set created for training; Section 3 discusses the architecture of the used model; Section 4 depicts the model training and its validation process, illustrating the results with some examples; Section 5 follows with an analysis of the obtained results. The document concludes with some final remarks and possible directions for future work.

## 2 The data-set

The Stanford Question Answering data-set (SQuAD) [8] contains a collection of around 400 generic articles from Wikipedia and, for each article, a set of questions and corresponding answers written in natural language. The collection is organized as a sequence of paragraphs (contexts) and, for each one, a set of questions and their answers is available. The answers can be gathered from the context paragraph.

Given the original data-set is only available in English, and there is no similar data-set available for the Portuguese language[1], we bootstrapped a Portuguese data-set using Machine Translation.

The original SQuAD data-set was automatically translated using the Google Translate API. Although the translation is not at human level, given that the paragraphs, questions, and answers are relatively small (just a few words in the case of the answers, a single sentence

[1] Note that there are other translations from SQuAD that are based on our work. Nevertheless, this article is relevant as our resource is being used for other works, and because implicitly we are evaluating Machine Translation systems.

for the questions, and a few sentences in the case of the contexts), most of the translations are acceptable. The main problem, as we will discuss later, is the lack of coherence between different translations, depending on the sequence context.

Given that the goal of the model is to find patterns between the structure of the text and the formulated questions, grammatical accuracy, exceptions, inconsistencies and similar shortcomings usually associated with automatic translations should not have a big influence in the results.

The current version of the translated data-set is available online[2]. The data-set is divided in two parts: *train* and *dev*, that have the same format, and can be used for training and testing respectively.

## 3 The Model

Bidirectional Encoder Representations from Transformers (BERT) is a state-of-the-art model for many NLP tasks as, for instance, document classification, named entity recognition, or question-answering [3]. The model is implemented using a deep neural network with a set of different layers. It can be seen as the complement of two major networks: (i) a pre-trained transformer that acts as a language model, and (ii) a network with different outputs depending on the task at hand.

In the context of question-answering tasks, the input to the model is composed by the context (the snippet of text that contains the answer) and the information query (the question). The output of the model is the span of text, from the context, that contains the answer. This is represented by the starting and ending tokens of the span of the answer in the context:

$$\text{qaptnet} :: (\text{Context}, \text{Question}) \rightarrow (\text{Start}, \text{End})$$

## 4 Training and Validation

Google provides different flavors of pre-trained parameters for the BERT model, both as monolingual or multilingual, and with different depth size (with different amounts of hidden layers). To train QAPTNET we used one of the multilingual models, case sensitive and 12 hidden layers, as a bootstrap model. The model was then trained using the training data-set, for 2 epochs, using a batch size of 8. The optimizer used was the Adam algorithm with a fixed weight decay, and a learning rate of $3 \times 10^{-5}$ [6]. Implementation, training and testing of the model was performed done using the PyTorch-Transformers Python package[3].

The model was then tested with the validation data-set, scoring around 50% accuracy, i.e. it was able to correctly find the answer for half the questions. The final version of the model is available publicly[4], including a Python package companion that helps using it. The main reason for this result is that the model is not able to completely capture a pattern between the question and the corresponding answer. This may be due to some factors including: not enough data for training, or the model not being complex enough to capture the intended pattern.

The following examples (in Portuguese) illustrate the use of the final model. For example, given the following `context`:

[2] Available at `https://github.com/nunorc/squad-v1.1-pt`.

[3] Available at `https://pytorch.org/hub/huggingface_pytorch-transformers/`.

[4] Available at `https://github.com/nunorc/qaptnet`.

> *Arquitetonicamente, a escola tem um caráter católico. No topo da cúpula de ouro do edifício principal é uma estátua de ouro da Virgem Maria. Imediatamente em frente ao edifício principal e de frente para ele, é uma estátua de cobre de Cristo com os braços erguidos com a lenda Venite Ad Me Omnes. Ao lado do edifício principal é a Basílica do Sagrado Coração. Imediatamente atrás da basílica é a Gruta, um lugar mariano de oração e reflexão. É uma réplica da gruta em Lourdes, na França, onde a Virgem Maria supostamente apareceu a Santa Bernadette Soubirous em 1858. No final da unidade principal (e em uma linha direta que liga através de 3 estátuas e da Cúpula de Ouro), é um estátua de pedra simples e moderna de Maria.*

One could pose the following `question`:

> *A quem a Virgem Maria supostamente apareceu em 1858 em Lourdes, na França?*

The model result is:

```
>>> qaptnet.query(context = context, question = question)
'Santa Bernadette Soubirous'
```

This example uses the model Python package companion to simplify the use of the model, a complete example on how to use this package is available in [1].

## 5 Data-Set and Results

One of the main challenges of this work is the data-set scale. In fact, it was automatically translated and not manually corrected. This means that some translations are not at a trained human translator level, and also that some translations were performed differently whether the string appeared with context (in the context paragraph) or without context (in the short answer). A simple example if the answer to the question shown in the previous section. While “Santa Bernadette Soubirous” is the correct answer, it will be accounted as wrong, as when translating the answer the system used “Saint Bernadette Soubirous”.

As a simple exercise to evaluate the quality of the data-set, for each answer we checked whether it is present in the context. Performing this analysis in the training data-set, we found up that 27 710 answers were not present in the context, from a total amount of 87 599 answers (about 30% of the training set). Doing the same analysis in the validation data-set, 11 081 answers of the total amount of 34 726 was not present in the context paragraph (about 31%).

This analysis did not account for other problems, like the lack of quality of the translation or even the upper-case versus lower-case comparison of the answers with the context (as we had a case-sensitive model).

Nevertheless, after training, a 50% accuracy was obtained. If we account the 30% of problematic entries, that the system would never guess, from the total amount of 34 736 entries, only 23 645 entries could be correct. Doing the proportion, the system would get up to 70% accuracy accounting only the correct parts of the validation set.

## 6 Conclusion

This paper introduces QAPTNET a model for performing question answering tasks, i.e. given a context and a question find the span of text in the context that answers the question, for the Portuguese language. The final version of the model achieves around 50% accuracy on the development part of the data-set, which is interesting enough for a first exploratory attempt, given the complexity of the task, and the initial lack of data to train the model. The other major outcome of this work is the data-set, the first of its' kind, in Portuguese, publicly available, that other researches can use to explore different models, and be improved by the community.

Returning to the research question guiding this work, the final model is able to find a satisfying number of answers for question, context pairs written in Portuguese. The transfer learning approach was the main enabler for achieving these results with such a small number of steps and shortcomings in the data-set. Of course this is still a work in progress, some trends for future work include:

- Correcting the data-set, validating the translations and the existence of the answer in the contexts;
- Fine-tuning model the hyper-parameters: some parameters used during training can be tuned to achieve better results, e.g. learning rate and batch size, also increasing the model complexity might help capturing patterns in the data;
- The analysis of a similar approach using the 2.0 version of SQuAD data-set.

## References

1 N.R. Carvalho, 2019 (last accessed: 28- 08-2019). URL: `https://github.com/nunorc/qaptnet`.

2 Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint*, 2019. `arXiv:1901.02860`.

3 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, 2018. `arXiv:1810.04805`.

4 SurveyMonkey Audience Drift and Myclever Salesforce. The 2018 state of chatbots report. how chatbots are reshaping online experiences, 2019.

5 Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

6 Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint*, 2014. `arXiv:1412.6980`.

7 Emilio Soria Olivas. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques: Algorithms, Methods, and Techniques*. IGI Global, 2009.

8 Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint*, 2016. `arXiv:1606.05250`.

9 Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764, 2019.