Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5442–5449
Marseille, 11–16 May 2020

AIA-BDE: A Corpus of FAQs in Portuguese and their Variations

Hugo Gonçalo Oliveira[1,2], João Ferreira[1,2], José Santos[1,2], Pedro Fialho[3,4], Ricardo Rodrigues[1,5], Luísa Coheur[3,6], Ana Alves[1,7]
[1]CISUC, Universidade de Coimbra, Portugal
[2]DEI, FCTUC, Universidade de Coimbra, Portugal
[3]INESC-ID, Lisboa, Portugal
[4]Universidade de Évora, Portugal
[5]ESEC, Instituto Politécnico de Coimbra, Portugal
[6]Instituto Superior Técnico, Universidade de Lisboa, Portugal
[7]ISEC, Instituto Politécnico de Coimbra, Portugal
hroliv@dei.uc.pt, jdcoelho@student.dei.uc.pt, santos@student.dei.uc.pt, pedro.fialho@l2f.inesc-id.pt, rmanuel@dei.uc.pt, luisa.coheur@l2f.inesc-id.pt, ana@dei.uc.pt

Abstract

We present AIA-BDE, a corpus of 380 domain-oriented FAQs in Portuguese and their variations, i.e., paraphrases or entailed questions, created manually, by humans, or automatically, with Google Translate. Its aims to be used as a benchmark for FAQ retrieval and automatic question-answering, but may be useful in other contexts, such as the development of task-oriented dialogue systems, or models for natural language inference in an interrogative context. We also report on two experiments. Matching variations with their original questions was not trivial with a set of unsupervised baselines, especially for manually created variations. Besides high performances obtained with ELMo and BERT embeddings, an Information Retrieval system was surprisingly competitive when considering only the first hit. In the second experiment, text classifiers were trained with the original questions, and tested when assigning each variation to one of three possible sources, or assigning them as out-of-domain. Here, the difference between manual and automatic variations was not so significant.

Keywords: FAQ retrieval, corpora creation, paraphrases detection, textual entailment, dialogue systems, Portuguese Language Processing

1. Introduction

Question-answering (QA) dialogue systems should be able to handle different ways of formulating the same information need. Therefore, besides measuring performance on giving the right answers for a given question, their ability to match a given interaction with suitable questions in their knowledge base is often key, and should also be assessed. This is the case, for instance, of Frequently Asked Questions (FAQs) retrieval systems (Karan et al., 2013; Caputo et al., 2016), which retrieve relevant FAQs for interactions in natural language.

This paper presents AIA-BDE, a corpus of FAQs, in Portuguese, and variations of their questions, which are reformulations that paraphrase or are entailed by the original questions, created automatically, with Google Translate, or manually, by volunteers. AIA-BDE was developed in the scope of a project that aims to develop more intelligent systems for supporting automatic assistance to entrepreneurs, using natural language (*AIA: Apoio Inteligente a Empreendedores*). In this project, different information retrieval (IR) and QA systems are being developed for answering questions in the domain of economic activities and their practice in Portugal (e.g., Gonçalo Oliveira et al. (2019), Santos et al. (2020)). In order to compare the performance of such systems without each time resorting to expensive and time-consuming human judgements, we developed AIA-BDE, which covers FAQs on the previous domains and has in mind the development and automatic assessment of the previous systems.

Reformulated questions can be seen as a simulation of user queries. Their manual creation was also a time-consuming process but, once it has been created, the corpus can be used several times for measuring progress and comparing different approaches for matching user interactions with FAQs. Furthermore, once available, it may also be used by other researchers, in other projects. Despite its original goal, AIA-BDE may further be seen as a benchmark for assessing how other systems of this kind adapt to this domain, or even as the starting point for related tasks, such as Semantic Textual Similarity (Cer et al., 2017) or Natural Language Inference (Bowman et al., 2015) in a QA scenario. To the best of our knowledge, there is no corpus with domain questions and their paraphrases in Portuguese, so this is also a gap we aim to fill.

In Section 2. we overview some related work on the evaluation of dialogue systems and resources used for this purpose, with some focus on FAQ retrieval systems, and on other corpora for QA in Portuguese. In Section 3. we describe the creation of AIA-BDE and show some examples of FAQs and variations. Before concluding, in Section 4. we report on the results of two experiments using AIA-BDE, aiming to provide additional insights. First, we have applied some unsupervised baselines for matching variations with their original questions and comment on the obtained results. It confirmed that this is a challenging task, especially when considering the manually created variations, more creative and, sometimes, incomplete. Another curiosity is that, in this scenario, the performance of Lucene IR system was very close to semantic similarity computed directly with state-of-the-art contextual word embeddings, namely ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). In the second experiment, we trained classifiers for automatically identifying the source of each FAQ, and tested them with the variations. The main conclusion was that, using a SVM,

macro-F1 was greater than 80% for any type of variation, and not significantly different for different types.

2. Related Work

Data-driven dialogue systems can learn from corpora like the Ubuntu Dialogue corpus (Lowe et al., 2015), Open Subtitles (Tiedemann, 2009) dialogues, or Twitter conversations (Ritter et al., 2011). Such systems (e.g., Vinyals and Le (2015)) are challenging to assess because they do not have clearly-defined goals. Towards their automatic evaluation, word-overlap metrics borrowed from machine translation (e.g., BLEU, METEOR) have been used to compare responses by the system with ground-truth responses produced by humans (Sordoni et al., 2015; Li et al., 2016). However, it is easy to understand that, due to the variety of valid responses in an open-domain conversation, such measures are not adequate, and it has been confirmed that they correlate very weakly with human judgements (Liu et al., 2016).

An expensive alternative is to ask users to interact with the system and leave their feedback on how human and natural their conversation was. When it comes to task-oriented dialogue systems, users may also answer to what extent their task was successfully accomplished (Wen et al., 2017). In any case, subjects can be recruited via crowd-sourcing, but, in order to measure progress, this would have to be done for each update.

Single-turn question-answering (QA) systems are a specific kind of task-oriented dialogue systems that allow users to search for information using natural language. Such systems typically target factoid questions (Voorhees, 2008) and rely on Information Retrieval (IR) techniques (Kolomiyets and Moens, 2011), which means that they can also be assessed with IR-based measures, namely the precision and the recall of the given answers, according to a ground-truth.

FAQs are useful resources for task-oriented dialogue systems, and corpora for evaluating FAQ retrieval systems have been created in Croatian (Karan et al., 2013) and Italian (Caputo et al., 2016). For training and testing a FAQ retrieval system for Croatian, based on STS, Karan et al. (2013) collected 1,222 FAQs, including questions, answers, type of service, and category, crawled from the website of a Croatian mobile phone operator. Then, without looking at the collected FAQs, ten human annotators invented user queries that would be asked by real users of the target operator, as well as their paraphrases. Finally, for each query, the binary relevance was manually set for each FAQ retrieved by a set of standard retrieval models (keyword search, phrase search, TF.IDF, language modelling).

QA4FAQ (Caputo et al., 2016) is a shared task on Question Answering for FAQs in Italian, organised in the scope of EVALITA 2016, with the goal of retrieving relevant FAQs for a given user query. It relied on 406 FAQs (question, answer, tags), 1,132 user queries collected from the logs of an IR system, and a set of mappings between queries and relevant FAQs.

Community Question Answering (Nakov et al., 2015; Nakov et al., 2016; Nakov et al., 2017) is a related task, with the goal of ranking question-comment and question-question similarity in web forums. As data source, its editions have used questions from the Qatar Living forum. Starting with a list of original questions, related questions were obtained for each, together with the first comments in their threads. Relevance was annotated for related questions to the original question, and for comments, to related questions and to the original question, though different annotations were used in different subtasks. Queries were then generated from the subject of each original question and Google used for collecting up to 200 question-comment threads in the forum site. Results with ten or more comments and questions with less than 2,000 characters were considered to be related questions.

Recently, datasets for question-answering within a dialogue context have also been developed, such as Choi et al. (2018) or Reddy et al. (2019), created both in a task where one person asks questions on a subject and the other answers, as naturally as possible, based on a text on the target subject.

Specifically concerning Portuguese, corpora of subtitles have been used for conversational agents, to better deal with out-of-vocabulary interactions (Magarreiro et al., 2014); there are collections of factoid questions (Santos and Rocha, 2004; Magnini et al., 2004) and topics (Mota et al., 2012) with their answers, previously used in IR and QA shared tasks, and dense domain questions (Criscuolo et al., 2017), also with their answers; as well as collections with pairs of sentences with their semantic similarity and entailment label (Fonseca et al., 2016; Real et al., 2020). Yet, as far as we know, there is no corpus with Portuguese domain-oriented questions and their variations, ready to be used in the evaluation of IR / QA dialogue systems.

3. Corpus Creation

The starting point of the AIA-BDE corpus were several groups of FAQs associated to services of the former *Balcão do Empreendedor* (BDE), the Portuguese Entrepreneur's Desk, now integrated in the e-Portugal website[1]. The BDE is a single point of access to digital services related to the exercise of economic activity in Portugal. It is directed to entrepreneurs who wish to perform services and obtain information inherent to the economic activities that they practice. More precisely, we collected 380 FAQs grouped in the following sources, which correspond, roughly, to available services:

- 118 FAQs from the Guide for the Application of the Legal Regime for Access and Exercise of Trade, Services and Catering Activities (*Guia de Aplicação do Regime Jurídico de Acesso e Exercício de Atividades de Comércio, Serviços e Restauração – RJACSR*);
- 56 FAQs from the Legislation of the Local Accommodation (*Legislação do Alojamento Local* – AL);
- 206 FAQs from the Business Spot (former *Portal Empresa*, now *Espaço Empresa* – PE), which targets the creation and management of businesses.

This means that we have three distinct groups of FAQs: RJACSR, AL and PE.

[1] `https://eportugal.gov.pt/`

Bearing in mind that, in most cases, users do not search for the exact question found in the FAQ, for each original question, we have added reformulations of such questions, i.e., different ways of asking the same questions, hereafter, variations. Those variations were created with two distinct approaches, namely:

- Using the Google Translate API[2] as a quick low-cost approach for generating paraphrases of the original questions, as follows: translation of Portuguese to English and back to Portuguese (VG1); the previous result back to English and to Portuguese again (VG2).
- Manual creation by a group of native Portuguese speaking volunteers, which were asked to provide alternative ways of asking each question (e.g., using different words / word-order), also considering that their answer would still to be valid. These do not include only paraphrases of the original questions, but also closely-related / entailed questions, possibly written in a more creative way, and a minority with spelling mistakes (VUC).

For each original question, there is a single VG1 and a single VG2 variation, but there are at least two VUC variations, with some questions having more. Therefore, the total of VG1 and VG2 variations is 380 of each kind, whereas for VUC there are 936 variations. Table 1 illustrates the AIA-BDE corpus with three FAQs, one of each covered source, with the original question (P) and answer (R) together with the FAQ variations of each kind, all with a rough English translation.

Due to their creation approach, the surface form of VG1 and VG2 tends to be very close to the original questions, often changing only one or two words. It is thus not strange to find a minority of VG1 that are exactly the same as the original questions, and some VG2 that are exactly the same as the VG1 for the same question. In addition, due to the simplicity of their creation, some of the resulting changes may end up changing slightly the semantics of the original question. On the other hand, VUC variations attempted at being more natural and are, at least, more creative. In this case, some variations are not exact paraphrases of the original questions, but are entailed by them. This also suggests that VUC variations should be harder to match with the corresponding original questions.

4. Experimentation with AIA-BDE

This section reports the results of two simple experiments on AIA-BDE, willing to provide some insights on the data and set some baselines for possible applications of the corpus. In the first experiment, several unsupervised approaches were used for matching variations with their original question. The second experiment targeted the automatic identification of the source of a variation.

4.1. Matching Variations to Original Questions

To better understand the challenge underlying the matching of variations with their original questions, we applied the following unsupervised approaches for this purpose:

- Using the **Chatterbot**[3] library that generates responses to a given input;
- Using the popular full-text search library **Lucene**[4];
- Semantic similarity computed from vector representations of text, obtained with the following pre-trained models:
  - FastText (Bojanowski et al., 2017) word embeddings, pre-trained for Portuguese[5];
  - ELMo (Peters et al., 2018) contextual word embeddings, pre-trained for Portuguese[6] (Gardner et al., 2018);
  - BERT (Devlin et al., 2019) contextual word embeddings, pre-trained for multiple languages[7].

The application of Chatterbot followed two simple steps: (i) training with all the original questions and answers; (ii) checking the responses given for each variation. Since the default implementation only provides an answer for a given interaction, we only compute its accuracy for the first hit.

For the remaining approaches, we also compute accuracy for the presence of the correct answer in the top-3 or top-5 best ranked candidates. This has in mind that, in many scenarios, it is better to return a smaller set of answers that include the correct one, than to give no answer or show one that is incorrect.

For using Lucene, the original questions were first indexed with the available PortugueseAnalyzer, which includes a stemming feature. After this, each variation was used as a search query on the created index, with candidate questions ranked by the default similarity scoring, namely BM25, with $k_1 = 1.2$ and $b = 0.75$. In order to compute accuracy on the first hit, top-3 and top-5, five ranked questions were retrieved for each query, and the correct question must be the first, in the top-3 or in the top-5, respectively.

When using any model of embeddings, each sentence was represented by a fixed-length vector of numbers, and similarity was computed with the cosine between the vector representation of each variation and the vector representation of all original questions. Accuracy is obtained from the number of variations for which the correct question was the most similar. For the top-3 and top-5, the correct question must be in the top-3 and top-5 most similar, respectively. The main difference is in how the vector of each question is computed.

For FastText and ELMo, each question and variation were first tokenized with the tokenizer of the NLTK toolkit, with additional rules for Portuguese clitics and contractions (Ferreira et al., 2019). With FastText, the sentence vector results from averaging the vectors of each token used, ignoring tokens without alpha-numeric characters (e.g., punctuation

[2] https://cloud.google.com/translate/docs/

[3] https://chatterbot.readthedocs.io/

[4] https://lucene.apache.org/

[5] https://fasttext.cc/

[6] https://allennlp.org/elmo

[7] https://github.com/google-research/bert/blob/master/multilingual.md

Source	Var	Text
RJACSR	P	*Qual a coima aplicável às contraordenações graves?* (What is the fine for serious offences?)
	VG1	*Qual é a multa aplicável à falta grave?* (What is the fine applicable to serious misconduct?)
	VG2	*Qual é a multa aplicável à falta grave?* (What is the fine applicable to serious misconduct?)
	VUC	*coima para contraordenação grave* (fine for serious offence)
	VUC	*Qual o valor da multa para contraordenações graves?* (How much is the fine for serious offences?)
	R	*As contraordenações graves são sancionáveis com coima: ...* (Serious offences are punishable with a fine ...)
AL	P	*No alojamento local é obrigatória a certificação energética? Em que termos deve ser efetuada?* Is energy certification compulsory in local accommodation? In what terms should it be done?
	VG1	*No alojamento local é obrigatório a certificação energética? Em que condições deveria ser feito?* (Is energy certification compulsory in local accommodation? Under what conditions should it be done?)
	VG2	*A certificação energética é necessária em alojamento local? Em que condições deve ser feito?* (Is energy certification required in local accommodation? Under what conditions should it be done?)
	VUC	*Como deve ser feita certificação energética do meu alojamento local?* (How should the energy certification of my local accomodation be done?)
	VUC	*Qual o procedimento para certificar energeticamente o meu alojamento local?* (What is the procedure to energetically certify my local accommodation?)
	R	*De acordo com esclarecimento da DGEG (Direção-Geral de Energia e Geologia) ...* (According to DGEG (General-Direction of Energy and Geology) ...)
PE	P	*Se a inscrição for recusada devolvem-me o valor do emolumento pago?* (If my registration is refused, will I be refunded for the paid emoluments?)
	VG1	*Se o pedido for recusado, serei reembolsado o valor da taxa paga?* (If the order is declined, will I be refunded for the amount of the paid fee?)
	VG2	*Se o pedido for rejeitado, serei reembolsado o valor da taxa paga?* (If the request is rejected, will I be refunded for the paid fee?)
	VUC	*Se a minha inscrição for rejeitada sou reembolsado?* (If my application is rejected am I refunded?)
	VUC	*No caso de uma inscrição ser recusada, o valor pago é devolvido?* (If an application is declined, is the paid amount sent back?)
	R	*O emolumento pago é devolvido, decorrido o prazo de 30 dias ...* (The paid emoluments are returned within 30 days ...)

Table 1: Examples of the AIA-BDE corpus.

signs) and tokens not covered by the model. For ELMo, a single embedding is assigned to the tokenized sentence.

BERT was used with the bert-as-a-service framework[8], which simplifies the process of obtaining a sentence embedding from BERT. To obtain Portuguese embeddings, we employ the model BERT-Base, Multilingual Cased, as provided and recommended by the BERT authors[9]. BERT models are based on WordPiece tokenization, where a token may include part of multiple words and not comply with typical word level segmentation (Devlin et al., 2019). Hence, with BERT we employed its default tokenizer. Moreover, we did not set a maximum on the sequence length.

We should add that both BERT and ELMo output multiple layers, each containing token embeddings on a certain depth of analysis. Yet, as suggested in the bert-as-a-service framework, we used the second last layer in both. We also used the default pooling strategy of bert-as-a-service, which means that, as it happened for FastText, the BERT embedding of each sentence was given by the arithmetic means of the (contextualized) word embedding vectors, in this case, of the second last layer. Since these embeddings were not fine-tuned, we did not use the [CLS] vector for this purpose.

Table 2 reports on the accuracy with each matching approach in AIA-BDE, measured by the ratio of variations for which the correct original question is the first hit, in the top-3 or top-5 hits. For validation purposes, the original questions (P) are also used as variations. And, although Lucene performed quite well in this scenario, it failed to match the following question with itself: *Qual a coima aplicável às contraordenações graves?* This question was instead matched with a similar one with an additional word: *Qual a coima aplicável às contraordenações muito graves?* As expected, accuracies are slightly better for VG1 than for VG2,

[8] https://github.com/hanxiao/bert-as-service/

[9] https://github.com/google-research/bert

and significantly lower for VUC. For both VG1 and VG2, accuracy on the first hit was higher than 90% for Lucene. ELMo was slightly better in VG1, and slightly worse in VG2. BERT was the worst of the three, but still with accuracies of 87% and 85%, respectively for the first hit in VG1 and VG2. The main difference is that, when looking at the top-3 or top-5, improvements are more pronounced for the contextual word embeddings. So much that Lucene is outperformed by both models of contextual embeddings, with ELMo achieving accuracies higher than 95% for both VG1 and VG2. When it comes to the manually-produced variations (VUC), behavior was similar, but accuracies lower. Though with only 67% accuracy, Lucene achieved the best results in the first hit again. Yet, for the presence in the top-3 and top-5, improvement is again more pronounced for the contextual embeddings, enough for ELMo achieving the best results, respectively close to 70% and 75% accuracy.

Method	Top	P	VG1	VG2	VUC
Chatterbot	1	98.4%	12.6%	9.2%	2.7%
Lucene	1	99.7%	90.3%	90.0%	**67.1%**
	3	100.0%	91.8%	91.3%	**68.7%**
	5	100.0%	91.8%	91.3%	68.8%
FastText	1	**100.0%**	65.3%	63.2%	31.0%
	3	100.0%	79.5%	76.3%	42.3%
	5	100.0%	82.9%	80.3%	48.7%
ELMo	1	**100.0%**	**91.8%**	**90.5%**	51.3%
	3	100.0%	**98.2%**	**97.9%**	67.3%
	5	100.0%	**98.7%**	**98.7%**	**72.6%**
BERT	1	**100.0%**	87.1%	85.0%	44.1%
	3	100.0%	93.9%	93.4%	57.9%
	5	100.0%	95.5%	94.7%	62.6%

Table 2: Accuracy of several baseline approaches when matching variations with the correct question.

Anyway, we stress that the reported results should be seen as mere baselines. Improvements would most certainly be obtained if the contextual models were fine-tuned with the original questions, or models were trained with a broader range of handcrafted features, also including those resulting from some of the previous approaches. For Portuguese, training could be done in the ASSIN collections (Fonseca et al., 2016; Real et al., 2020), where pairs of sentences have an assigned value of similarity, between 1 (completely different) and 5 (equivalent). Though, at the same time, it might not be the most suitable data, because most sentences are in the declarative form, whereas AIA-BDE contains interrogative sentences. Differences on detecting semantic similarity between the latter have been studied (Rodrigues et al., 2018). This also supports the fact that, in the future, AIA-BDE could potentially be used for training Natural Language Inference (NLI) systems that identify paraphrases of interrogative sentences.

4.2. Classifying the Source of Variations

This second part of the work was a text classification experiment, where classifiers were trained with the original questions, using the source of the FAQs as a label. We thus had three classes: RJACSR, AL and PE. Yet, bearing in mind that a user might also input out-of-domain (OOD) interactions, for which there is no available answer, we added a fourth class, OOD, with a random selection of 206 questions – i.e., interactions ending with a '?' – from a corpus of movie subtitles in Portuguese (Magarreiro et al., 2014). The number 206 was chosen to be the same as the number of FAQs with the most frequent source, PE.

The ability to perform this classification might be useful for different scenarios. For instance, given an interaction, it might avoid computing the similarity with every known question, and thus improve time complexity. Or it could be used as plan B in a single-turn QA dialogue system: when no questions can be retrieved, it may at least indicate a related service, and possibly suggest the user to go to some document or website about it. Alternatively, it may identify the interaction as out-of-domain and opt for producing a response with a different strategy.

Classifiers were learned with three different methods, namely a Linear SVM, a Naïve Bayes (NB) and a Random Forest (RF) Classifier, all implemented in the scikit-learn (Pedregosa et al., 2011) library. In all three methods, questions were represented by their TF-IDF-weighted vectors with up to 200 bag-of-words features, considering only words that occurred in 50% or less training questions. Additional parameters of each method were the scikit-learn defaults.

The learned classifiers were then used in three datasets, one with each set of variations, in order to automatically assign their source. However, variations were mixed with another random selection of questions from movie subtitles, in the same quantity as the number of variations from the PE source, 206 for VG1 and VG2, and 434 for VUC. Table 3 presents the results obtained with this experiment.

Though not always for specific classes, the best Macro-F1 was always achieved with the SVM, which suggests that this should be the method to use for this purpose. In opposition to the previous experiment, here, performance differences between the automatically produced and the manual variations was not so clear. We can say that, despite resulting from more changes in the original questions, they can still be identified as being on the same domain as their original questions. We recall that this might be useful when no question is matched.

On the best results, F1 was slightly above 80% for each type of variation. According to other experiments, if the OOD interactions were not used, this proportion would increase between 1 (VG2) and 6% (VUC). Especially for the VUC variations, for which matching with the original question was worse, identifying the correct service with about 80% precision might work as a plan B.

A complementary experiment was to approach this text classification task with the best method, SVM, but using BERT (Devlin et al., 2019) for encoding the text in a 768-dimension vector, in a similar fashion to what was done in the experiment of section 4.1.. As table 4 shows, minor improvements were achieved this way, still without fine-tuning BERT.

Again, we stress that better results could probably be obtained after the optimisation of some parameters in the classifiers, or if classifiers were dropped and a fine-tuned version of BERT was used directly for classification. For this reason, the presented results should be seen as mere baselines.

Method	Source	VG1			VG2			VUC		
		P	R	F1	P	R	F1	P	R	F1
SVM	RJACSR	69%	**88%**	**77%**	68%	**88%**	**77%**	**78%**	**71%**	**74%**
	AL	**98%**	77%	**86%**	**100%**	68%	81%	**88%**	69%	**77%**
	PE	**91%**	77%	**83%**	**90%**	76%	**82%**	**85%**	85%	**85%**
	OOD	87%	**91%**	**89%**	85%	**91%**	**88%**	80%	**91%**	**86%**
	Macro-Avg	**86%**	**83%**	**84%**	**86%**	**81%**	**82%**	**83%**	**79%**	**81%**
NB	RJACSR	68%	85%	76%	68%	85%	75%	**78%**	69%	73%
	AL	59%	**88%**	71%	59%	**86%**	70%	51%	**89%**	65%
	PE	77%	81%	79%	77%	**80%**	78%	71%	**86%**	78%
	OOD	**90%**	61%	73%	**88%**	61%	72%	**89%**	60%	72%
	Macro-Avg	74%	79%	75%	73%	78%	74%	72%	76%	72%
RF	RJACSR	**71%**	81%	75%	**69%**	81%	75%	70%	58%	63%
	AL	88%	82%	85%	90%	80%	**85%**	69%	61%	64%
	PE	83%	**83%**	**83%**	82%	78%	80%	72%	79%	75%
	OOD	89%	84%	86%	86%	84%	85%	78%	84%	81%
	Macro-Avg	83%	82%	82%	82%	**81%**	81%	72%	70%	71%

Table 3: Results when classifying different types of variation and out-of-domain interactions into their original source.

Method	Source	VG1			VG2			VUC		
		P	R	F1	P	R	F1	P	R	F1
SVM	RJACSR	**86%**	66%	75%	**86%**	64%	**83%**	**88%**	**74%**	**80%**
	AL	86%	**91%**	**89%**	86%	**91%**	**89%**	77%	**70%**	74%
	PE	85%	**87%**	**86%**	83%	**85%**	**84%**	**90%**	**85%**	**87%**
	OOD	**90%**	**99%**	**94%**	**90%**	99%	**94%**	**81%**	**99%**	**89%**
	Macro-Avg	**87%**	**86%**	**86%**	**86%**	**85%**	**85%**	**84%**	**82%**	**83%**

Table 4: Results when classifying different types of variation and out-of-domain interactions into their original source using BERT for representation.

5. Concluding Remarks and Future Work

We have described the AIA-BDE corpus, with FAQs on a specific domain, in Portuguese, and their reformulations (variations), some produced automatically and others manually. This corpus can be used as a benchmark for assessing FAQ retrieval systems, in Portuguese, other related, such as dialogue systems, or, indirectly, models of STS or NLI.

AIA-BDE is publicly available in a specific Github repository[10] and can now be used by other researchers in some of the aforementioned tasks, hopefully contributing to advance their state-of-the-art. It is organised in a simple text file where each line starts with a tag that indicates the kind of the following variation – P: for the original questions, R: for their answers, and VG1, VG2, VUC for the variations of the immediately previous questions.

We have also reported on two experiments where several unsupervised approaches were used for matching variations with the appropriate question, and where models were trained for automatically identifying the source of a specific variation.

Results of the first confirmed that manually created variations are hard to match with the original question. A somehow surprising result is that, when considering only the first hit, a traditional IR-based approach performed really close to state-of-the-art contextual word embeddings, better for two types of variation. Yet, when considering the top-3 and top-5 hits, contextual word embeddings were closer, with ELMo achieving the best accuracy, without any kind of fine-tuning. The better performance of ELMo against BERT might be explained by the utilization of a ELMo model pre-trained for Portuguese, while BERT was a multilingual model.

On the classification experiment, performance differences when of assigning the correct source to different kinds of variation were not so clear. This suggests that, despite resulting from more changes in the original questions, they can still be identified as being on the same domain, which might be useful when no question is matched.

Besides continuing to use this corpus as a benchmark, in a near future, we plan to, at least, triple its size with more FAQs from the new e-Portugal website, covering many more sources. In fact, a new set of 675 FAQs, recently obtained, is already in the AIA-BDE repository. Yet, we have not integrated it with the original corpus because additional effort is needed for dealing with the granularity of services / sources and with a minority of question overlap issues, which do not occur only at the surface level. Automatic variations for these were already generated with Google Translate. On the other hand, the creation of manual variations is time-consuming and we might have to resort to a crowdsourcing platform. We may also explore alternative methods for automated paraphrase generation (Barreiro, 2009).

Acknowledgements

This work was supported by Fundação para a Ciência e Tecnologia's (FCT), though the INCoDe 2030 initiative, in the scope of the demonstration project AIA, "Apoio Inteligente a empreendedores (chatbots)", which also supports the scholarships of João Ferreira, José Santos and Pedro

[10] https://github.com/hgoliv/AIA-BDE

Fialho; and by national funds through FCT under project UIDB/50021/2020.

6. Bibliographical References

Barreiro, A. (2009). Make it simple with paraphrases: Automated paraphrasing for authoring aids and machine translation. Ph.D. thesis, Universidade do Porto.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

Caputo, A., Degemmis, M., Lops, P., Lovecchio, F., and Manzari, V. (2016). Overview of the EVALITA 2016 question answering for frequently asked questions (QA4FAQ) task. In Proc 3rd Italian Conference on Computational Linguistics (CLiC-it 2016) & 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), volume 1749 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). SemEval-2017 task 1: Semantic Textual Similarity multilingual and crosslingual focused evaluation. In Procs. of 11th Intl. Workshop on Semantic Evaluation (SemEval-2017), pages 1–14. Association for Computational Linguistics.

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W.-t., Choi, Y., Liang, P., and Zettlemoyer, L. (2018). QuAC: Question answering in context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2174–2184, Brussels, Belgium, October-November. Association for Computational Linguistics.

Criscuolo, M., Fonseca, E. R., Aluísio, S. M., and Sperança-Criscuolo, A. C. (2017). MilkQA: a dataset of consumer questions for the task of answer selection. In Proceedings of the 6th Brazilian Conference on Intelligent Systems (BRACIS), volume 1, pages 354–359, Uberlândia, Brazil, October. IEEE.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

Ferreira, J., Gonçalo Oliveira, H., and Rodrigues, R. (2019). Improving NLTK for processing Portuguese. In Symposium on Languages, Applications and Technologies (SLATE 2019), page In press, June.

Fonseca, E., Santos, L., Criscuolo, M., and Aluísio, S. (2016). Visão geral da avaliação de similaridade semântica e inferência textual. *Linguamática*, 8(2):3–13.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., and Zettlemoyer, L. (2018). AllenNLP: A deep semantic natural language processing platform. In Proceedings of Workshop for NLP Open Source Software (NLP-OSS), pages 1–6, Melbourne, Australia, July. Association for Computational Linguistics.

Gonçalo Oliveira, H., Filipe, R., Rodrigues, R., and Alves, A. (2019). Using Lucene for developing question-answering agent in Portuguese. In Proceedings of 8th Symposium on Languages, Applications and Technologies (SLATE 2019), volume 74 of *OASIcs*, pages 2:1–2:14. Schloss Dagstuhl, June.

Karan, M., Žmak, L., and Šnajder, J. (2013). 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. Association for Computational Linguistics.

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

Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., and Dolan, B. (2016). A persona-based neural conversation model. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 994–1003, Berlin, Germany, August. Association for Computational Linguistics.

Liu, C.-W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2122–2132, Austin, Texas, November. Association for Computational Linguistics.

Lowe, R., Pow, N., Serban, I., and Pineau, J. (2015). The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 285–294, Prague, Czech Republic, September. Association for Computational Linguistics.

Magarreiro, D., Coheur, L., and Melour, F. S. (2014). Using subtitles to deal with out-of-domain interactions. In Proc 18th Workshop on the Semantics and Pragmatics of Dialogue (SemDial), pages 98–106.

Magnini, B., Vallin, A., Ayache, C., Erbach, G., Peñas, A., de Rijke, M., Rocha, P., Simov, K. I., and Sutcliffe, R. F. E. (2004). Overview of the CLEF 2004 Multilingual Question Answering track. In Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum (CLEF), Revised Selected Papers, volume 3491 of *LNCS*, pages 371–391. Springer.

Mota, C., Simões, A., Freitas, C., Costa, L., and Santos, D. (2012). Págico: Evaluating Wikipedia-based information retrieval in Portuguese. In Proceedings of the

8th International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May. ELRA.

Nakov, P., Màrquez, L., Magdy, W., Moschitti, A., Glass, J., and Randeree, B. (2015). SemEval-2015 task 3: Answer selection in community question answering. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 269–281, Denver, Colorado, June. Association for Computational Linguistics.

Nakov, P., Màrquez, L., Moschitti, A., Magdy, W., Mubarak, H., Freihat, A. A., Glass, J., and Randeree, B. (2016). SemEval-2016 task 3: Community question answering. In Proc 10th International Workshop on Semantic Evaluation. Association for Computational Linguistics, June.

Nakov, P., Hoogeveen, D., Màrquez, L., Moschitti, A., Mubarak, H., Baldwin, T., and Verspoor, K. (2017). SemEval-2017 task 3: Community question answering. In Proc 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 27–48. Association for Computational Linguistics, August.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (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, June. Association for Computational Linguistics.

Real, L., Fonseca, E., and Oliveira, H. G. (2020). 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.

Reddy, S., Chen, D., and Manning, C. D. (2019). Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Ritter, A., Cherry, C., and Dolan, W. B. (2011). Data-driven response generation in social media. In Proceedings of the conference on empirical methods in natural language processing, pages 583–593. Association for Computational Linguistics.

Rodrigues, J., Saedi, C., Branco, A., and Silva, J. (2018). Semantic equivalence detection: Are interrogatives harder than declaratives? In Proc 11th Language Resources and Evaluation Conference, Miyazaki, Japan, May. ELRA.

Santos, D. and Rocha, P. (2004). The key to the first CLEF with Portuguese: topics, questions and answers in CHAVE. In Workshop of the Cross-Language Evaluation Forum for European Languages, pages 821–832. Springer.

Santos, J., Alves, A., and Gonçalo Oliveira, H. (2020). Leveraging on Semantic Textual Similarity for developing a Portuguese dialogue system. In Computational Processing of the Portuguese Language - 13th International Conference, PROPOR 2020, Évora, Portugal, March 2-4, 2020, Proceedings, volume 12037 of *LNCS*, pages 131–142. Springer.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., and Dolan, B. (2015). A neural network approach to context-sensitive generation of conversational responses. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 196–205, Denver, Colorado, May–June. Association for Computational Linguistics.

Tiedemann, J. (2009). News from opus-a collection of multilingual parallel corpora with tools and interfaces. In Recent advances in natural language processing, volume 5, pages 237–248.

Vinyals, O. and Le, Q. V. (2015). A neural conversational model. In Proceedings of ICML 2015 Deep Learning Workshop, Lille, France.

Voorhees, E. M. (2008). Evaluating question answering system performance. In *Advances in Open Domain Question Answering*. Springer, pp. 409–430.

Wen, T.-H., Vandyke, D., Mrkšić, N., Gašić, M., Rojas-Barahona, L. M., Su, P.-H., Ultes, S., and Young, S. (2017). A network-based end-to-end trainable task-oriented dialogue system. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 438–449, Valencia, Spain, April. Association for Computational Linguistics.