Question Answering for Dialogue State Tracking in Portuguese

Francisco Pais, Patrícia Ferreira, Catarina Silva
CISUC, DEI, LASI
University of Coimbra, Portugal

Ana Alves
CISUC, ISEC, LASI
Polytechnic Institute of Coimbra, Portugal

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

fmpais@student.dei.uc.pt, {patriciaf,catarina,ana,hroliv}@dei.uc.pt

Abstract

Dialogue State Tracking (DST) is a component of task-oriented dialogue systems, used to track the progress of a conversation while maintaining a representation of the current state. We explore DST in Portuguese dialogues, marking the first known application specific to this language. We introduce a new task-oriented dialogue dataset in Portuguese, adapted from the widely-used MultiWOZ, and propose to leverage available question-answering (QA) models for slot filling. Predefined questions are made to user's utterance, in a process that does not require training in dialogue data. We evaluate two QA models, based on BERT-base and on T5, select suitable thresholds on their scores, and test both intent recognition, as a preliminary step, and post-processing for matching categorical slots. Performance is still far from the state-of-the-art for English, but incorporating intent recognition and post-processing significantly improves performance. These findings not only advance DST within Portuguese-speaking communities but also create opportunities for new dialogue systems in Portuguese.

Keywords: Dialogue Systems; Dialogue State Tracking; MultiWOZ; Slot-filling; Question-Answering; Intent Recognition.

1 Introduction

More and more people use dialogue systems for everyday tasks. These vary from simple actions like checking the weather to more intricate operations that require transactions and more computational processing, such as booking the cheapest flight to a specific location at a particular time.

Despite extensive research, many methodologies employed in dialogue system development exhibit limitations. One strategy is centered on creating agents through manual work (Zue et al., 2000; Wang and Lemon, 2013; Sun et al., 2014). This involves designing dialogue flows, defining relevant entities, and identifying potential intentions using phrases or keywords. An alternative strategy emphasizes the automatic generation of responses based on collections of human dialogues (Vinyals and Le, 2015; Zhang et al., 2020). Despite the lower manual effort and straightforward adaptation, earlier systems stemming from this strategy frequently displayed repetition and inconsistency (Williams, 2014; Henderson et al., 2013), leading to challenges in critical applications such as customer support.

Moreover, the aforementioned strategies struggle with context, often neglecting previously posed questions and being unable to leverage relationships between questions and answers in the same conversation; or they represent context in embeddings that are not interpretable by humans, thus not ready for a manual inspection. This is a compelling motivation for seeking methods that adeptly handle context by manipulating human-readable structures. Notably, most research in these domains is in English. For Portuguese, task-oriented dialogue (TOD) datasets in Portuguese that could assist in evaluating context monitoring are not freely available (e.g., (Xu et al., 2020)) or are the result of machine translation (e.g., (Ding et al., 2021)).

To address the challenge of context monitoring, we employed Dialogue State Tracking (DST) (Williams et al., 2016), an integral part of the Dialogue State architecture. DST keeps track of the state of an ongoing dialogue with a "slot filling" mechanism that fills specific slots based on the user's most recent actions within the conversation. To our knowledge, this is the first time DST has been applied with a focus on the Portuguese language.

Given the parallelism between slot value extraction and extractive question answering (QA), we propose to leverage models fine-tuned for this task. For Portuguese, both BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020) have been fine-tuned in the SQuAD (Rajpurkar et al., 2016) dataset, in order to answer open-domain natural language questions based on a given context. This is a cheaper alternative to training in an annotated dialogue dataset, often unavailable.

The proposed approach was experimented in MultiWOZ-PT (Ferreira et al., 2024), a recent adaptation of the widely-used MultiWOZ (Budzianowski et al., 2018) dataset whose utterances were translated to Portuguese and the database was adapted to a Portuguese city. Questions were predefined for each slot, and promising results were obtained after: (i) narrowing down the questions made with intent recognition; (ii) selecting suitable thresholds on the model confidence, for increased precision; (iii) integrating a post-processing step, for increased recall. Reported performance sets a baseline for future work, which will become more accessible with the release of MultiWOZ-PT.

The remainder of the paper is organized as follows: Section 2 discusses related work on DST and dialogue analysis in Portuguese; Section 3 presents the proposed approach; Section 4 reports on experimental results and on evaluation; Section 5 concludes the paper and discusses directions for future work.

2 Related Work

Earlier work on DST was driven by the Dialogue State Tracking challenges (Williams et al., 2016), but modern DST relies on two main types of neural approach: span-based, where slot values are extracted directly from the input utterances; and slot value generation. In both approaches, the BERT (Devlin et al., 2019) language model has been used for encoding the dialogue context. SUMBT (Lee et al., 2019) learns slot-value relationships through an attention mechanism; and BERT-DST (Chao and Lane, 2019) predicts slot values through classification heads, but this is done independently for each turn, instead of considering the full dialogue history.

Span-based approaches may also formulate DST as a reading comprehension task (Gao et al., 2019). Here, dialogue is seen as a context to which a natural language question is asked regarding the dialogue state (DS, e.g., *what is the value for slot x?*). Similarly to extractive question answering (QA), this question is to be answered with spans of the given context.

Value-generation approaches, such as TRADE (Wu et al., 2019) and MA-DST (Kumar et al., 2020), include an attention-based copy mechanism for capturing the correlation between slots and history, then generating a DS. SOM-DST (Kim et al., 2019) is similar, but, for predicting whether a slot needs to be updated, it takes both the previous dialogue turn and the previous DS as input.

The main limitation of span-based approaches is that the slot value is not always found exactly in the text. On the other hand, generation approaches tend to produce invalid values.

Hybrid approaches try to reduce the impact of the previous limitations. DS-DST (Zhang et al., 2019) adopts a dual strategy where the answers for categorical slots are selected from the possible values, and answers for non-categorical slots are extracted from the context with a reading comprehension model. TripPy (Heck et al., 2020) considers three types of slot values and adopts a different copy strategy for getting each. Values explicitly expressed by the user are extracted with a span-based approach; values expressed by the system and referred by the user are extracted from the system inform memory; values expressed earlier in the dialogue, i.e., co-references, are extracted from the dialogue history.

Most of the previous approaches were assessed, for English, in the MultiWOZ dataset (Budzianowski et al., 2018), primarily using joint goal accuracy (JGA) as the metric. JGA quantifies the proportion of dialogue turns for which the prediction, encompassing all slot-value pairs, is correct, i.e., matches the ground-truth dialogue state. Reported values for JGA in MultiWOZ 2.1 are between 42% (Lee et al., 2019) and 55% (Heck et al., 2020).

The approach adopted in this paper can be seen as hybrid in the sense that it extracts slot values from the user utterance (span-based), but then post-processes the values of categorical slots, in order to map them to valid ones. DST is also seen as a reading comprehension task, but an available QA model, not trained for this task, is repurposed for slot filling. Previously, QA models were used for Information Extraction with some success (Ferreira et al., 2023).

Joint intent recognition and slot filling were previously attempted in Portuguese, with a multilingual approach on MultiATIS (Xu et al., 2020), a proprietary dialogue dataset. It relied on a multilingual BERT encoder and explored machine translation and label projection methods for multilingual training and cross-lingual transfer. On the other hand, we tackle DST specifically for Portuguese and rely on the recent translation of a part of MultiWOZ to this language, which we make publicly available.

Early work on dialogue analysis and applications for Portuguese includes an approach for parsing multiple data types in dialogue systems, relying on expectations for better recognising objects in user utterances (Martins et al., 2008); or Natural Language Understanding (NLU) as a classification task with SVMs (Mota et al., 2012).

More recently, a conversational assistant was developed for smart homes (Ketsmur et al., 2019), with the NLU component delegated to IBM Watson. Still in the scope of NLU, embeddings and clustering were explored for automating the annotation of entities and intents in a dataset of (Covid-related) conversations (Júnior et al., 2021).

Other dialogue-related tasks applied to Portuguese include response generation for conversational agents (Melo and Coheur, 2020), learned from a small character-specific corpus and from a corpus of movie subtitles; or sentiment analysis on customer-support conversations (Carvalho et al., 2022).

3 Proposed Approach

In this section, we outline our approach for DST in Portuguese, which aims at facilitating slot-filling tasks, enabling enhanced context monitoring and, consequently, improved interactions. The section starts with an overview of the proposed approach, followed by its instantiation to our scenario, where we detail the dataset used, models for intent recognition and QA, and post-processing methods.

3.1 Overview

Figure 1 depicts the pipeline we employ for DST. The process starts with an utterance, generally by the user, which may be followed by intent recognition, in order to restrict slot filling to slots related to the target intent. After this, QA models are applied for slot filling. When data is scarce for training a model for this task, as it happens for Portuguese, we propose to use models for extractive QA trained in open-domain questions. Given a context (in this case, the utterance) and a question, these models extract a suitable answer from the context and provide a score on their confidence.

In order to ignore answers with lower confidence, thus increasing precision, we may consider only answers with confidence above a predefined threshold. In the proposed pipeline, this can be useful for ignoring slots that are not mentioned in the utterance. This is especially common when using models that were not trained for DST and always provide an answer to a question.

The final step is also optional and targets only categorical slots. Since the utterances may not refer the slot values verbatim, post-processing methods can be applied for mapping the user text to valid values. Figure 2 has a running example of the proposed pipeline, where an utterance goes through each step to finally fill a slot.

3.2 MultiWOZ-PT

MultiWOZ (Budzianowski et al., 2018) is a task-oriented dialogue (TOD) dataset, encompassing 10,000 dialogues with multiple interactions between two human participants: one assuming the role of a user, who has a task to accomplish; the other acting as the system, aiming to promptly respond to the user's requests, assisting in task completion. The utterances of this dataset cover multiple domains and are labelled with intents, slots, and their values.

Since MultiWOZ is in English, it cannot be used for training and testing dialogue systems in other languages. Together with the lack of a publicly available dataset of this kind for Portuguese, this motivated the manual adaptation of (a portion of) MultiWOZ to this language. While a Portuguese version of this dataset exists within the GlobalWOZ collection (Ding et al., 2021), it is important to note that it is the result of machine translation. Upon examining the samples of the corpus[1], it becomes evident that the quality of the machine-translated dialogues is poor. This is due to the brevity of utterances, the frequent presence of named entities, and the crucial role of context.

MultiWOZ-PT (Ferreira et al., 2024) is based on the test portion of MultiWOZ 2.2 (Zang et al., 2020), but its utterances are manually translated to Portuguese and its database is adapted to the city of Coimbra, Portugal, instead of Cambridge, UK. Being known by its old university, Coimbra ends up sharing some similarities with Cambridge. Information on the services of Coimbra was primarily obtained from well-known platforms, such as TripAdvisor[2] (mainly for restaurants), Book-

[1] See https://github.com/bosheng2020/globalwoz

[2] https://www.tripadvisor.pt/

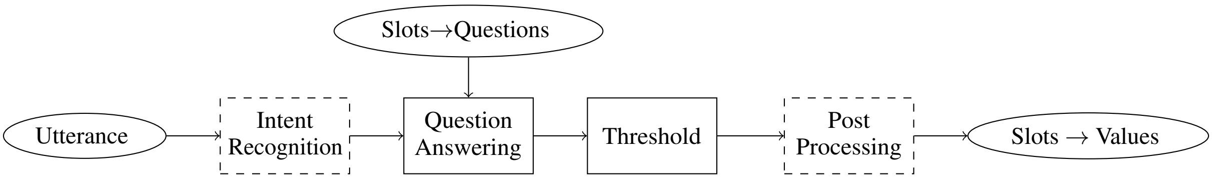
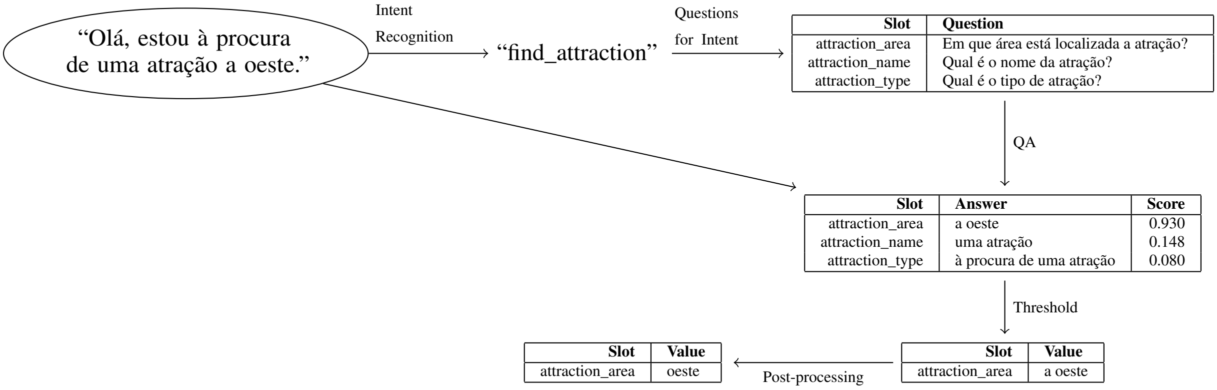


Figure 1: Pipeline for the slot–filling Approach. Dotted boxes represent optional steps in the pipeline and ellipses represent inputs and outputs.



Slot	Question
attraction_area	Em que área está localizada a atração?
attraction_name	Qual é o nome da atração?
attraction_type	Qual é o tipo de atração?

Slot	Answer	Score
attraction_area	a oeste	0.930
attraction_name	uma atração	0.148
attraction_type	à procura de uma atração	0.080

Slot	Value
attraction_area	a oeste

Slot	Value
attraction_area	oeste

Figure 2: Running example of the proposed approach.

ing[3] (for hotels), or CP[4] (for trains). A dictionary for mapping the original predefined values of categorical slots, in English, to Portuguese, is also provided.

MultiWOZ-PT contains 1,000 dialogues, originally divided in two files: one with 512 dialogues, the other with 488. These corresponded to the test dialogues of MultiWOZ, which cover five domains (restaurant, attraction, hotel, taxi, train) and 30 slot types. Out of its 1,928 utterances, 399 are related to attractions, 396 to hotels, 445 to restaurants, 198 to taxis, and 490 to trains. During translation, semantic consistency was maintained for preserving intents (find or book), domains, and slot values. Slots can be categorical, with possible values limited to a predefined set (e.g., in the hotel domain, valid values for "type" and "pricerange" are respectively "guesthouse" or "hotel", and "expensive", "cheap" or "moderate"); or non-categorical, with open values (e.g., the "address" slot). No slot types were introduced beyond those in the original dataset. However, the values of non-categorical slots were adapted to reflect the services in Coimbra.

Table 1 illustrates a complete dialogue from MultiWOZ, in English, and its adaptation to Portuguese in MultiWOZ-PT. For each user utterance, it includes information on intents, slots, and their values. MultiWOZ-PT was made publicly available[5], hopefully contributing to improving the state of the art of Portuguese dialogue systems.

3.3 Considering Intents

In the context of dialogue systems, intent recognition is the task of identifying the underlying goal of an utterance. Once this intent is recognized, the system can handle the utterance appropriately, e.g., by generating a response that fulfills the request. The utterances of MultiWOZ-PT are labelled with one of eight intent categories: find_attraction, find_hotel, book_hotel, find_restaurant, book_restaurant, find_taxi, find_train and book_train. For instance, the intent of the utterance "Eu gostaria de encontrar um hotel em Coimbra" (*I would like to find a hotel in Coimbra*) is "find_hotel". Since different intents have different slot types associated, knowing the

[3] https://www.booking.com/index.pt-pt.html

[4] https://www.cp.pt/passageiros/pt

[5] See https://github.com/NLP-CISUC/MultiWOZpt

Speaker	MultiWOZ 2.2	MultiWOZ-PT
USER	I need info on a train that would be departing from Peterborough. **Intent:** find_train **Slots:** "train-departure": "Peterborough"	Preciso de informações sobre um comboio que parta da Figueira da Foz. **Intent:** find_train **Slots:** "train-departure": "Figueira da Foz"
SYS	What day and time?	A que dia e hora?
USER	I would like to leave on Sunday and arrive in Cambridge by 15:15. **Intent:** find_train **Slots:** "train-arriveby": "15:15", "train-day": "Sunday", "train-departure": "Peterborough", "train-destination": "Cambridge"	Gostaria de partir no domingo e chegar a Coimbra pelas 15:15. **Intent:** find_train **Slots:** "train-arriveby": "15:15", "train-day": "Sunday", "train-departure": "Figueira da Foz", "train-destination": "Coimbra"
SYS	I have train TR7864 leaving at 14:19 and arriving at 15:09. Would you like to book that?	Tenho o comboio 16819 com partida às 13:58 e chegada às 15:09. Gostaria de o reservar?
USER	That'd be perfect, I need three tickets on Sunday. **Intent:** book_train **Slots:** "train-arriveby": "15:15", "train-bookpeople": "3", "train-day": "Sunday", "train-departure": "Peterborough", "train-destination": "Cambridge"	Isso seria perfeito, preciso de três bilhetes para domingo. **Intent:** book_train **Slots:** "train-arriveby": "15:15", "train-bookpeople": "3", "train-day": "Sunday", "train-departure": "Figueira da Foz", "train-destination": "Coimbra"
SYS	Booking was successful, the total fee is 39.59 GBP payable at the station. Reference number is: HS2FFYBI. Is there anything else I can assist you with today?	A reserva foi efetuada com sucesso, a taxa total é de 8,40 euros a pagar na estação. O número de referência é: HS2FFYBI. Posso ajudá-lo com mais alguma coisa hoje?
USER	I'm also looking for information on a place called the man on the moon. **Intent:** find_attraction **Slots:** "attraction-name": "the man on the moon"	Também estou à procura de informações sobre um lugar chamado Salão Brazil. **Intent:** find_attraction **Slots:** "attraction-name": "Salão Brazil"
SYS	The man on the moon is a concert hall in the center. Would you like to know anything else?	O Salão Brazil é uma sala de concertos no centro. Gostaria de saber mais alguma coisa?
USER	Yes, can I get the address to that please? **Intent:** find_attraction **Slots:** "attraction-name": "the man on the moon"	Sim, pode dar-me o endereço, por favor. **Intent:** find_attraction **Slots:** "attraction-name": "Salão Brazil"
SYS	Yes, the address is 2 Norfolk Street.	Sim, o endereço é Largo do Poço, nº3, 1º Andar.
USER	Can you tell me what the area of town is? Can I get the phone number too, please? **Intent:** find_attraction **Slots:** "attraction-name": "the man on the moon"	Pode dizer-me qual é a zona da cidade? Pode também dar-me o número de telefone, por favor? **Intent:** find_attraction **Slots:** "attraction-name": "Salão Brazil"
SYS	It's located in the center of town. The phone number is 01223474144. Do you need assistance with anything else?	Situa-se no centro da cidade. O número de telefone é 239837078. Precisa de ajuda com mais alguma coisa?
USER	That will be it for today! Thank you so much!	É tudo por hoje! Muito obrigado!
SYS	You're very welcome! Have a great day!	Não tem de quê! Tenha um ótimo dia!

Table 1: Original dialogue ID PMUL1241 from MultiWOZ and its translation in MultiWOZ-PT with intents, slots, and slot values for each user utterance.

intent may help in narrowing the slots to extract, hopefully reducing noise and increasing precision.

For experimentation, the intent annotations of MultiWOZ-PT can be used. However, in the real world, the intent of each utterance would have to be automatically recognized. As MultiWOZ-PT is a new dataset, there are no models available for this. So, we fine-tuned two available language models for intent recognition in MultiWOZ-PT: BERTimbau-base (Souza et al., 2020), based on BERT (Devlin et al., 2019); and Albertina-PTPT (Rodrigues et al., 2023), based on DeBERTa (He et al., 2020). Both models were used through the transformers library and the Hugging-Face hub[6][7]. Details of the training process can be found in Section 4.1.

3.4 QA for Slot Filling

Towards slot filling, the proposed approach leverages available models for QA. Adopting QA models that are available off-the-shelf is a cheaper alternative to training a model specific for DST, for which available data would not be enough. While MultiWOZ-PT contains only eight intents, which enabled the training of a classifier, the number of different slots amounts to 30, but still only 1,000 dialogues, out of which some have to be held out

[6] BERTimbau available from https://huggingface.co/neuralmind/bert-base-portuguese-cased

[7] Albertina available from https://huggingface.co/PORTULAN/albertina-900m-portuguese-ptpt-encoder

for testing.

In order to handle different types of question, as well as language variability, it is important to use models trained in a large number of contexts and questions, produced by different annotators. The SQuAD dataset (Rajpurkar et al., 2016) features 100,000 question-answer pairs crafted by crowd-workers, based on given contexts from Wikipedia articles, thus covering multiple domains. Every question is answered with a passage from the context. For Portuguese, there are transformer-based models fine-tuned in a translation of SQuAD to Portuguese, based on known architectures like BERT (Devlin et al., 2019) or T5 (Raffel et al., 2020). The main difference between the previous is that the BERT models extract the answer directly from the context, whereas T5 is a text-to-text model that generates the answers.

Following this, the models explored in our work are both available from the HuggingFace hub[8][9] and resulted from fine-tuning BERTimbau (Souza et al., 2020) and PTT5 (Carmo et al., 2020). A fine-tuned model is available for each version of BERTimbau, base and large, but, after noticing that differences between both were minimal, we decided to use only the smaller BERTimbau-base.

In order to get the slot values from an utterance, questions are made to the selected models, using the utterance as context[10]. This required the formulation of 30 natural language questions, one for each slot in MultiWOZ-PT. Questions were hand-crafted, but this has to be done only once for each dataset / inventory of slots. Questions were the result of preliminary tests, where we tried to follow a similar style as in SQuAD, making questions as straightforward as possible, and always mentioning the name of the slot.

Table 2 illustrates the questions used with those related to the hotel domain. The same questions were used for both models, BERT and T5. The full list of questions for all the slots is revealed in Appendix A.

Slot Type	Question
hotel-area	*Em que área está localizado o estabelecimento?*
hotel-bookday	*Em que dia é a reserva?*
hotel-bookpeople	*Quantas pessoas são?*
hotel-bookstay	*Quantos dias vai ficar?*
hotel-internet	*Tem internet grátis?*
hotel-name	*Qual é o nome do estabelecimento?*
hotel-parking	*Tem estacionamento gratuito?*
hotel-pricerange	*Qual é o preço médio do estabelecimento?*
hotel-stars	*Quantas estrelas tem?*
hotel-type	*Qual é o tipo de estabelecimento?*

Table 2: Questions for the Hotel Domain.

3.5 Post-Processing

The user will not always mention the slot value verbatim in the utterance. In some cases, the expected value will be inflected (e.g., plural instead of singular). In other cases, the model will give an answer that is longer than the slot value.

In order to increase recall, we try to match the given answers with valid values. This is, however, only possible for categorical slots, which have a known fixed set of valid values.

So, two methods were adopted for matching the answer with the closest slot value: the Levenshtein Distance (Lev) and Semantic Textual Similarity (STS). Lev is an established method for measuring the distance between two strings as the number of editions necessary for transforming one into the other. In opposition to Lev, which is language agnostic and does not consider semantics, STS computes the cosine of sentence embeddings. For this, we relied on a sentence transformer available from HuggingFace[11], based on BERTimbau fine-tuned in sentence pairs from shared tasks on semantic similarity (Fonseca et al., 2016; Real et al., 2020). Table 3 illustrates the application of the post-processing methods with real examples.

Method	Answer	Matched with
Lev	arquitetónico	arquitetura
	zona este	este
	residenciais	residencial
	cara	caro
STS	depois do meio dia	meio dia
	sexta-feira às 16:00	sexta-feira
	chinês	chinesa
	centro da cidade	centro

Table 3: Examples of answers correctly matched with valid slot values, using Lev and STS for post-processing.

4 Experimentation

This section reports on the experimentation of the proposed approach in MultiWOZ-PT and its evalua-

[8] BERTimbau-base for QA: https://huggingface.co/pierreguillou/bert-base-cased-squad-v1.1-portuguese

[9] PTT5 for QA: https://huggingface.co/pierreguillou/t5-base-qa-squad-v1.1-portuguese

[10] Using a variable number of previous utterances as context was also tested, but it generally resulted in lower performance.

[11] https://huggingface.co/rufimelo/bert-large-portuguese-cased-sts

tion. It includes the evaluation of intent recognition, the selection of thresholds, and the evaluation of DST. Since MultiWOZ-PT is divided in two files, for a more natural split, we used the first file, which contains 512 dialogues, as our training set, and the remaining 488 dialogues for testing.

4.1 Evaluation of Intent Recognition

Towards their incorporation in the proposed approach, the selected models (see Section 3.3) were fine-tuned for intent recognition. In this process, the following hyperparameters were used for both BERTimbau and Albertina: batch size 32, learning rate of $1e^{-5}$, and training duration of 5 epochs.

Table 4 reports on their precision (P), recall (R) and F1-Score (F1) when they are fine-tuned in the training dialogues and evaluated in the test.

Intent	**Albertina-PTPT**			**BERTimbau**		
	P	**R**	**F1**	**P**	**R**	**F1**
find_attract	0.76	0.90	0.83	0.81	0.90	0.85
find_hotel	0.80	0.83	0.81	0.81	0.84	0.82
book_hotel	0.72	0.76	0.74	0.78	0.75	0.76
find_rest	0.84	0.77	0.80	0.87	0.75	0.81
book_rest	0.78	0.88	0.83	0.72	0.84	0.78
find_taxi	0.95	0.74	0.83	0.80	0.82	0.81
find_train	0.92	0.89	0.90	0.91	0.88	0.89
book_train	0.83	0.63	0.72	0.79	0.75	0.77
Macro Avg	0.82	0.80	0.81	0.81	0.82	0.81
Weight Avg	0.83	0.83	0.83	0.83	0.83	0.83

Table 4: Intent Recognition performance for MultiWOZ-PT: Performance (P), Recall (R), and F1-Score.

Both models show similar performance overall, with precision, recall, and F1-Score exceeding 0.80. They perform better for intents like "find_train" ($F1 \approx 0.90$), followed by "find_attraction" and "find_hotel," which are the three most represented in the dataset, and perform less effectively for "book_hotel" and "book_train," the least represented ones.

Since using one or the other model would not make much difference, we decided to use BERTimbau in the following experiments, because it is a more established model with much fewer parameters (110M vs 900M).

4.2 Dialogue State Tracking

Even though the proposed approach does not use models trained in the DST task, the thresholds applied to the confidence of the QA models can be optimized. Before reporting on the performance of the models used, this section reports the threshold optimisation step, performed in the 512 training dialogues of MultiWOZ-PT.

4.2.1 Thresholds Optimisation

Threshold optimisation consisted of assessing the proposed approach with a range of threshold values for finally selecting the best performing for each slot. After some preliminary tests, the following ranges were tested in the 512 training dialogues: $[0.49, 0.59, 0.69, 0.79]$ for the BERT model; $[0.80, 0.85, 0.90, 0.95]$ for the T5 model. The performance of DST with these thresholds was computed both without and with post-processing.

The selection of the optimal thresholds was guided by plots like those in Figure 3. For the two QA models, these plots depict the evolution of precision for the slots of the restaurant domain. Optimal values are marked with a red star. A table with all the values selected for each slot, QA model and post-processing method is in Appendix B. Those were the values used in the following experiments.

4.2.2 Evaluation of DST

After selecting the optimal thresholds, the proposed approach was applied to the 488 test dialogues of MultiWOZ-PT, and metrics commonly used for assessing DST were computed. The Joint Goal Accuracy (JGA) quantifies the proportion of dialogue turns for which the prediction, encompassing all slot-value pairs, is correct. Slot F1 evaluates DST on a per-slot basis, by computing the harmonic mean of the system precision (i.e., the proportion of accurate slot-value predictions out of all slot-value predictions by the system) and recall (i.e., the proportion of accurate slot-value predictions out of all true slot values in the dialogue) for each slot in the dialogue. Whereas the JGA is more strict, and expects the system to be accurate in all aspects of the dialogue state, Slot F1 assesses the system performance for individual slots.

Tables 5 and 6 report on the evaluation of each model considering three different approaches for handling intents: (i) Intent=None means that intents were not considered, i.e., the QA model tries to get a value for each of the 30 slots; (ii) Intent=Gold, where the intent recognition is based on the labels of the dataset; (iii) Intent=Classifier, where the intent recognition is based in the fine-tuned BERTimbau model described in Section 4.1. Scores are presented for the slots of each domain and overall.

We first observe that, regardless of the variations in intent recognition and post-processing, the T5 model is always outperformed by the BERT model. JGA is far from perfect and always lower than Slot

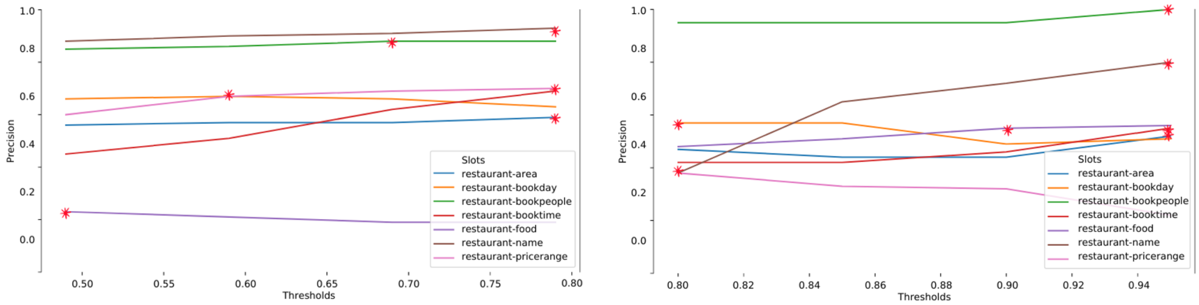


Figure 3: Optimizing thresholds for the slots of the Restaurant domain, using Levenshtein for post-processing. On the left, for the BERT model, and on the right, for the T5 model.

	Intent = None						**Intent = Gold**						**Intent = Classifier**					
	JGA			**Slot F1**			**JGA**			**Slot F1**			**JGA**			**Slot F1**		
Domain	None	Lev	STS	None	Lev	STS	None	Lev	STS	None	Lev	STS	None	Lev	STS	None	Lev	STS
Attraction	0.08	0.17	0.15	0.18	0.33	0.27	0.23	0.35	0.34	0.42	0.51	0.49	0.25	0.36	**0.37**	0.46	**0.53**	0.52
Hotel	0.09	0.12	0.09	0.21	0.27	0.20	0.25	**0.30**	0.25	0.46	0.50	0.43	0.27	0.29	0.27	0.50	**0.52**	0.45
Restaurant	0.08	0.17	0.10	0.24	0.37	0.27	0.20	0.22	**0.23**	0.49	0.49	0.50	0.19	0.22	0.20	0.52	**0.54**	0.50
Taxi	0.04	0.08	0.15	0.08	0.16	0.33	0.30	0.34	0.35	0.41	0.48	0.45	0.32	0.35	**0.39**	0.47	**0.51**	0.50
Train	0.14	0.34	0.14	0.40	0.52	0.36	0.28	0.48	0.32	0.63	0.70	0.60	0.30	**0.51**	0.32	0.65	**0.72**	0.56
Weight Avg	0.10	0.20	0.12	0.26	0.36	0.28	0.25	**0.34**	0.29	0.51	0.55	0.50	0.26	0.32	0.29	0.54	**0.58**	0.50

Table 5: Performance of the BERT-base model for each domain, considering different intent recognition and post-processing methods.

	Intent = None						**Intent = Gold**						**Intent = Classifier**					
	JGA			**Slot F1**			**JGA**			**Slot F1**			**JGA**			**Slot F1**		
Domain	None	Lev	STS	None	Lev	STS	None	Lev	STS	None	Lev	STS	None	Lev	STS	None	Lev	STS
Attraction	0.04	0.10	0.11	0.11	0.21	0.22	0.13	0.23	0.23	0.37	0.49	0.49	0.14	**0.25**	**0.25**	0.38	0.51	**0.52**
Hotel	0.07	0.09	0.07	0.17	0.21	0.16	0.21	0.23	0.21	0.38	0.41	0.37	0.23	**0.25**	0.23	0.41	**0.44**	0.39
Restaurant	0.05	0.11	0.07	0.18	0.26	0.23	0.15	**0.22**	0.17	0.40	0.50	0.45	0.12	0.20	0.14	0.43	**0.52**	0.46
Taxi	0.03	0.05	0.13	0.08	0.13	0.29	0.22	0.22	0.26	0.38	0.40	0.44	0.26	0.26	**0.32**	0.45	0.46	**0.49**
Train	0.11	0.35	0.12	0.37	0.50	0.34	0.23	0.50	0.28	0.62	**0.72**	0.59	0.24	**0.51**	0.27	0.63	**0.72**	0.56
Weight Avg	0.07	0.17	0.10	0.21	0.30	0.24	0.19	0.30	0.23	0.45	**0.53**	0.47	0.19	**0.31**	0.23	0.48	0.51	0.48

Table 6: Performance of the T5-base model for each domain, considering different intent recognition and post-processing methods.

F1 scores. This was expected, since JGA is a strict measure.

Still, performance improves when intents are considered, no matter where they come from. This confirms that, by narrowing down the target slots, intent recognition is a critical step for DST. Additionally, we note that differences between using the gold intents or those by an automatic classifier are minimal. In fact, the best average Slot F1 (0.58) was achieved with the intent classifier.

Post-processing has also a positive impact. Here, the Levenshtein distance is particularly noteworthy, as it is always the best option overall and for most domains. Despite being limited to string editions, it is possible that only a small fraction of answers actually diverge from the target value in more than a few characters (e.g., synonyms), i.e., where STS would be preferable.

Despite the low performance, the best JGA (0.34) still means that, for more than one third of the dialogue turns, all slots were correctly filled. As the first approach to DST in MultiWOZ-PT, we see this as promising, though with room for future improvements.

5 Conclusion

In this paper, we proposed an approach for Dialogue State Tracking (DST) that leverages available models for Question Answering (QA) and experimented it in dialogues in Portuguese. We have shown that, when training data is scarce, these models can be seen as an alternative to slot filling.

Despite the low Joint Goal Accuracy (JGA), we have shown that performance can improve significantly if: slots are narrowed down by intent recognition; the model confidence is considered and suitable thresholds are applied; and the values for categorical slots are post-processed.

The best JGA (0.34) is achieved by BERTimbau fine-tuned for QA, leveraging the intents in the dataset, and Levenshtein for post-processing. It is still far from reference scores for English (i.e., between 0.42 for span-based and 0.55 for hybrid approaches), but, in any case, it means that more than one third of the dialogue turns have all their slots correctly filled.

We remind that, to the best of our knowledge, this is the first work on DST focused on Portuguese, which was only possible after the adaptation of the MultiWOZ TOD dataset to this language. So, there is definitely room for future improvements. In the context of the proposed approach, alternative questions (e.g., obtained through prompt engineering) and post-processing methods may be tested (e.g., BLEU (Papineni et al., 2002)).

We could also consider state-of-the-art DST methods, such as those referred to in Section 2. Since all of them are supervised in DST, the main obstacle remains to be the availability of enough dialogues with annotated intents and slots. On this scope, we will consider augmenting MultiWOZ-PT by translating more dialogues of the original dataset, following the same guidelines.

Acknowledgements

This work was supported by: the project POWER (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); the Portuguese Recovery and Resilience Plan through project C645008882-00000055, Center for Responsible AI; and by national funds through FCT, within the scope of the project CISUC (UID/CEC/00326/2020).

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. 2020. PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data. *arXiv preprint arXiv:2008.09144*.

Isabel Carvalho, Hugo Gonçalo Oliveira, and Catarina Silva. 2022. Sentiment Analysis in Portuguese Dialogues. In *Proceedings of IberSPEECH 2022*, pages 176–180. ISCA.

Guan-Lin Chao and Ian Lane. 2019. BERT-DST: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. *arXiv preprint arXiv:1907.03040*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of 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. ACL Press.

Bosheng Ding, Junjie Hu, Lidong Bing, Sharifah Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. Globalwoz: Globalizing multiwoz to develop multilingual task-oriented dialogue systems. *arXiv preprint arXiv:2110.07679*.

Bruno Carlos Luís Ferreira, Hugo Gonçalo Oliveira, and Catarina Silva. 2023. Leveraging question answering for domain-agnostic information extraction. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Proceedings of 26th Iberoamerican Congress on Pattern Recognition (CIARP)*, volume 14469 of *LNCS*, pages 244–256. Springer.

Patrícia Ferreira, Francisco Pais, Catarina Silva, Ana Alves, and Hugo Gonçalo Oliveira. 2024. MultiWOZ-PT: A task-oriented dialogue dataset in Portuguese. Submitted to LREC-COLING 2024.

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

Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tur. 2019. Dialog state tracking: A neural reading comprehension approach. *arXiv preprint arXiv:1908.01946*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gašić. 2020. Trippy: A triple copy strategy for value independent neural dialog state tracking. *arXiv preprint arXiv:2005.02877*.

Matthew Henderson, Blaise Thomson, and Steve Young. 2013. Deep neural network approach for the dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 467–471.

Valmir Oliveira Dos Santos Júnior, Joao Araújo Castelo Branco, Marcos Antonio De Oliveira, Ticiana L Coelho Da Silva, Lívia Almada Cruz, and Regis Pires Magalhaes. 2021. A natural language understanding model Covid-19 based for chatbots. In *2021 IEEE 21st International conference on bioinformatics and bioengineering (BIBE)*, pages 1–7. IEEE.

Maksym Ketsmur, António Teixeira, Nuno Almeida, and Samuel Silva. 2019. Towards European Portuguese Conversational Assistants for Smart Homes. In *Proceedings of the 8th Symposium on Languages, Applications and Technologies (SLATE 2019)*, volume 74 of *OpenAccess Series in Informatics (OASIcs)*, pages 5:1–5:14. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee. 2019. Efficient dialogue state tracking by selectively overwriting memory. *arXiv preprint arXiv:1911.03906*.

Adarsh Kumar, Peter Ku, Anuj Goyal, Angeliki Metallinou, and Dilek Hakkani-Tur. 2020. Ma-DST: Multi-attention-based scalable dialog state tracking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8107–8114.

Hwaran Lee, Jinsik Lee, and Tae-Yoon Kim. 2019. SUMBT: Slot-utterance matching for universal and scalable belief tracking. *arXiv preprint arXiv:1907.07421*.

Filipe M Martins, Ana Mendes, Joana Paulo Pardal, Nuno J Mamede, and Joao P Neto. 2008. Using system expectations to manage user interactions. In *Computational Processing of the Portuguese Language: 8th International Conference, PROPOR 2008 Aveiro, Portugal, September 8-10, 2008 Proceedings 8*, pages 240–243. Springer.

Gonçalo Melo and Luísa Coheur. 2020. Towards a conversational agent with "character". In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings 14*, pages 420–424. Springer.

Pedro Mota, Luísa Coheur, Sérgio Curto, and Pedro Fialho. 2012. Natural language understanding: From laboratory predictions to real interactions. In *Text, Speech and Dialogue: 15th International Conference, TSD 2012, Brno, Czech Republic, September 3-7, 2012. Proceedings 15*, pages 640–647. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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

Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. 2020. The ASSIN 2 shared task: A quick overview. In *Computational Processing of the Portuguese Language - 14th International Conference, PROPOR 2020, Evora, Portugal, March 2-4, 2020, Proceedings*, volume 12037 of *LNCS*, pages 406–412. Springer.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Freitas Osório. 2023. Advancing neural encoding of Portuguese with transformer AlbertinaPT-*. In *Progress in Artificial Intelligence – 22nd EPIA Conference on Artificial Intelligence, EPIA 2023, Faial Island, Azores, September 5-8, 2023, Proceedings, Part I*, volume 14115 of *LNCS*, pages 441–453. Springer.

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

Kai Sun, Lu Chen, Su Zhu, and Kai Yu. 2014. A generalized rule based tracker for dialogue state tracking. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 330–335. IEEE.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. In *Proceedings of ICML 2015 Deep Learning Workshop*, Lille, France.

Zhuoran Wang and Oliver Lemon. 2013. A simple and generic belief tracking mechanism for the Dialog State Tracking challenge: On the believability of observed information. In *Proceedings of the SIGDIAL 2013 Conference*, pages 423–432.

Jason D Williams. 2014. Web-style ranking and SLU combination for dialog state tracking. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 282–291.

Jason D Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. *arXiv preprint arXiv:1905.08743*.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online. Association for Computational Linguistics.

Jian-Guo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wan, Philip S Yu, Richard Socher, and Caiming Xiong. 2019. Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. *arXiv preprint arXiv:1910.03544*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Victor Zue, Stephanie Seneff, James R Glass, Joseph Polifroni, Christine Pao, Timothy J Hazen, and Lee Hetherington. 2000. Juplter: a telephone-based conversational interface for weather information. *IEEE Transactions on speech and audio processing*, 8(1):85–96.

A Questions for each slot

Table 7 enumerates the questions handcrafted for extracting the values of each slot.

Slot Type	**Question**
attraction-area	*Em que área está localizada a atração?*
attraction-name	*Qual é o nome da atração?*
attraction-type	*Qual é o tipo de atração?*
hotel-area	*Em que área está localizado o estabelecimento?*
hotel-bookday	*Em que dia é a reserva?*
hotel-bookpeople	*Quantas pessoas são?*
hotel-bookstay	*Quantos dias vai ficar?*
hotel-internet	*Tem internet grátis?*
hotel-name	*Qual é o nome do estabelecimento?*
hotel-parking	*Tem estacionamento gratuito?*
hotel-pricerange	*Qual é o preço médio do estabelecimento?*
hotel-stars	*Quantas estrelas tem?*
hotel-type	*Qual é o tipo de estabelecimento?*
restaurant-area	*Em que área está localizado o restaurante?*
restaurant-bookday	*Em que dia é a reserva?*
restaurant-bookpeople	*Quantas pessoas são?*
restaurant-booktime	*A que horas é a reserva?*
restaurant-food	*Qual é tipo de comida?*
restaurant-name	*Qual é o nome do restaurante?*
restaurant-pricerange	*Qual é o preço médio do restaurante?*
taxi-arriveBy	*A que horas chega?*
taxi-departure	*De onde quer sair?*
taxi-destination	*Para onde quer ir?*
taxi-leaveAt	*A que horas é que sai?*
train-arriveBy	*A que horas chega?*
train-bookpeople	*Quantas pessoas são?*
train-day	*Em que dia é a reserva?*
train-departure	*De onde quer sair?*
train-destination	*Para onde quer ir?*
train-leaveAt	*A que horas é que sai?*

Table 7: Natural language questions handcrafted for each slot.

B Optimal Thresholds

Table 8 reports on the optimal thresholds selected for each slot, QA model, post-processing method. These slots, selected on 512 training dialogues, were used in the evaluation of DST.

	BERT-base			**T5-base**		
Slots	None	Lev	STS	None	Lev	STS
area	0.59	0.79	0.79	0.90	0.90	0.90
name	0.79	0.79	0.79	0.95	0.95	0.95
type	0.69	0.79	0.79	0.95	0.90	0.95
area	0.69	0.69	0.69	0.95	0.95	0.95
bookday	0.79	0.79	0.79	0.90	0.90	0.90
bookpeople	0.59	0.69	0.69	0.80	0.80	0.80
bookstay	0.49	0.49	0.49	0.95	0.95	0.95
internet	0.79	0.69	0.69	0.95	0.95	0.95
name	0.79	0.79	0.79	0.95	0.95	0.95
parking	0.79	0.49	0.49	0.95	0.95	0.95
pricerange	0.79	0.59	0.59	0.90	0.95	0.80
stars	0.79	0.79	0.69	0.95	0.95	0.95
type	0.69	0.79	0.69	0.95	0.95	0.95
area	0.59	0.79	0.69	0.95	0.95	0.95
bookday	0.49	0.59	0.49	0.80	0.80	0.85
bookpeople	0.69	0.69	0.69	0.95	0.95	0.95
booktime	0.79	0.79	0.69	0.95	0.95	0.90
food	0.69	0.49	0.79	0.90	0.90	0.95
name	0.79	0.79	0.69	0.95	0.95	0.95
pricerange	0.69	0.79	0.59	0.95	0.80	0.90
arriveBy	0.79	0.79	0.79	0.95	0.95	0.95
departure	0.79	0.79	0.79	0.95	0.95	0.95
destination	0.79	0.79	0.79	0.95	0.95	0.95
leaveAt	0.79	0.79	0.79	0.95	0.95	0.95
arriveBy	0.79	0.79	0.79	0.95	0.95	0.95
bookpeople	0.69	0.79	0.79	0.95	0.95	0.90
day	0.49	0.49	0.49	0.95	0.95	0.95
departure	0.79	0.79	0.79	0.95	0.95	0.95
destination	0.79	0.79	0.79	0.95	0.95	0.95
leaveAt	0.79	0.69	0.79	0.95	0.95	0.95

Table 8: Selection of Optimal Thresholds for different variations of the QA Models.