# Dialogue State Tracking in Task-Oriented Portuguese Dialogues

Francisco Pais
fmpais@student.dei.uc.pt
Patrícia Ferreira
patriciaf@dei.uc.pt
Catarina Silva
catarina@dei.uc.pt
Hugo Gonçalo Oliveira
hroliv@dei.uc.pt

Universidade de Coimbra
CISUC - Centro de Informática e Sistemas
FCTUC-DEI - Departamento de Engenharia Informática
Coimbra, Portugal

## Abstract

Dialogue State Tracking (DST) is a technique for monitoring the current state of a conversation by keeping track of filled slots and representing the user's most recent actions within the dialogue. In this paper, we explore an unsupervised approach based on Question-Answering (QA) for DST in task-oriented Portuguese dialogues. Given the absence of existing task-oriented datasets in Portuguese, we decided to create the first such dataset, inspired by the widely recognized MultiWOZ dataset. Among the models we tested, the best results were achieved using BERT-base and BERT-large, fine-tuned on the SQuAD dataset. The choice between the two will depend on the specific task the user wishes to solve. These models showed to be promising for task-oriented dialogue systems in Portuguese, especially when lacking training data.

## 1 Introduction

DST is crucial for dialogue systems, as it enables them to represent the context of a conversation and respond accordingly. In practical scenarios, such as conversations between a user and a dialogue system intended to accomplish specific tasks like booking a hotel or restaurant, DST enhances task completion by monitoring the entire dialogue. It uses a process called "slot-filling", which fills specific slots with information needed to complete the task. User-imposed constraints, like a hotel's location in the city center or a restaurant's specific cuisine, are also considered. We executed slot-filling using three QA models: BERT-base[1], BERT-large[2], and T5[3]. Additionally, we evaluated these QA models both with and without the incorporation of Intent Detection and Post-Processing methods.

The premise of this approach is that when dialogue systems employ DST, they not only execute tasks more accurately from the outset but also save the user from having to repeat information. This is a common issue in systems that do not monitor conversations, leading users to question the system's capability and reliability. Moreover, this lack of monitoring wastes user's time as they find themselves repeating previously information, a significant concern in today's fast-paced world.

In the remaining sections of this paper, we will cover several key areas. First, we will present work related to our research. Next, we will describe our approach, which required the creation of a task-oriented dataset. Using this dataset, we tested various models and evaluated their performance. Finally, we will highlight the main conclusions and discuss potential directions for future research.

## 2 Related Work

In this section, we focus on discussing algorithms utilized for DST within the well-known MultiWOZ dataset [1]. This is especially relevant given that we have applied the DST method to a Portuguese dataset, which has not been subjected to other algorithms thus far. The MultiWOZ dataset served as the foundation for creating the first version of our Portuguese dataset. We have named this dataset MultiWOZpt, and it is publicly available on GitHub[4].

Several supervised learning algorithms, which require training data, have been tested on the MultiWOZ dataset. Starting with TripPy [3], which employs various copying mechanisms to populate the slots with values. These values are extracted from the real-time dialogue context. Slots can be filled using one of three mechanisms: Span prediction, which extracts values directly from the user's sentence; copying a value from a system inform memory that keeps track of the system's inform operations; or copying a value from a different slot already present in the dialogue state (DS) to resolve coreferences within and across domains. Another algorithm, SUMBT [6], uses BERT to encode slot IDs and candidate values. It learns slot-value relationships in dialogues through an attention mechanism. BERT-DST [2] employs contextual representations to encode each dialogue turn, feeding them into classification heads for value prediction.

An alternative to span prediction is value generation. TRADE [8] and MA-DST [5] use a copy mechanism to generate a DST by combining distributions over a predefined vocabulary and the vocabulary of the current context. SOM-DST [4] also uses similar mechanisms for value generation but takes the previous dialogue turn and DS as input to BERT in order to predict the current DS. It incorporates a state operation predictor to determine whether a slot needs to be updated or not. However, generative models have the drawback of potentially producing invalid values, such as word repetitions or omissions. A hybrid approach known as DS-DST has been proposed to address this issue. It utilizes both span-based and picklist-based predictions for slot-filling [9]. Unlike generative approaches, picklist-based and span-based methods use existing word sequences to populate slots. DS-DST partially mitigates the limitations of span prediction by employing a picklist method to fill a subset of slots.

## 3 Dataset

Given that the primary objective was to explore effective methods for monitoring context, and considering that much of the existing research in this domain has been conducted in English, there were no readily available datasets of annotated dialogues in Portuguese for evaluating context monitoring. Consequently, it became necessary to develop the first task-oriented dataset in Portuguese, which we named MultiWOZpt. This new dataset was created by adapting and translating a small portion of the existing MultiWOZ dataset to fulfill our main objective.

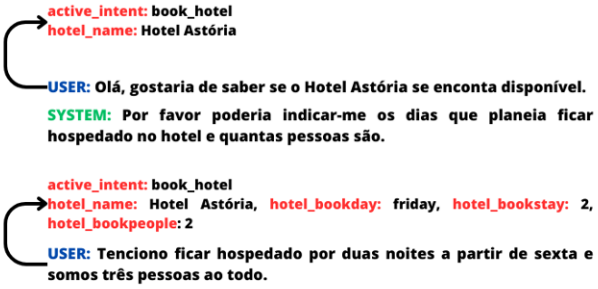
The dataset was co-created by two annotators. We began by manually translating and adapting 512 dialogues from the first test partition of the MultiWOZ dataset into Portuguese. During this process, we utilized a specially designed database to align services in Cambridge, England, with those in Coimbra, Portugal, the city of our university. For example, if a dialogue in the original dataset involved a user searching for an attraction, such as a museum in Cambridge, we sought equivalent museums in Coimbra as substitutes.

Upon completing this first version of the Portuguese dataset, we conducted an analysis to examine its composition, the results can be found in Table 1. An excerpt from a dialogue in the MultiWOZpt dataset can be seen in Figure 1, which highlights some of its features.

| Service | # Examples | # Slots per Service | # Intents |
|---|---|---|---|
| Attraction | 199 | 3 | 530 |
| Hotel | 201 | 10 | 779 |
| Restaurant | 237 | 7 | 781 |
| Taxi | 105 | 4 | 219 |
| Train | 261 | 6 | 934 |
| **Total** | 1,003 | 30 | 3,243 |

Table 1: Analysis of the Structure of the MultiWOZpt Dataset.

[1] https://huggingface.co/pierreguillou/bert-base-cased-squad-v1.1-portuguese
[2] https://huggingface.co/pierreguillou/bert-large-cased-squad-v1.1-portuguese
[3] https://huggingface.co/pierreguillou/t5-base-qa-squad-v1.1-portuguese
[4] https://github.com/NLP-CISUC/MultiWOZpt/


active_intent: book_hotel
hotel_name: Hotel Astória

USER: Olá, gostaria de saber se o Hotel Astória se enconta disponível.

SYSTEM: Por favor poderia indicar-me os dias que planeia ficar hospedado no hotel e quantas pessoas são.

active_intent: book_hotel
hotel_name: Hotel Astória, hotel_bookday: friday, hotel_bookstay: 2, hotel_bookpeople: 2

USER: Tenciono ficar hospedado por duas noites a partir de sexta e somos três pessoas ao todo.

Figure 1: Excerpt from a dialogue of the MultiWOZpt dataset.

# 4 Approach and Results

DST was employed to address the issue of insufficient context monitoring in dialogues. For this reason, we leveraged on available models fine-tuned for QA and used them for slot-filling. Given a natural language question, these models are specifically designed to identify relevant sequences within a text or corpus. Their ability to accurately extract information from user utterances to specific slots is noteworthy. Equally important is their ability to handle variability. Users can express the same intent in multiple ways, making it important for these models to be trained on a diverse array of questions and contexts to accommodate this variability. This allows them to correctly identify slots even when users express intentions differently. The models we used were trained on a Portuguese-translated version of the SQuAD[7] dataset. Figure 2 illustrates how these models work.

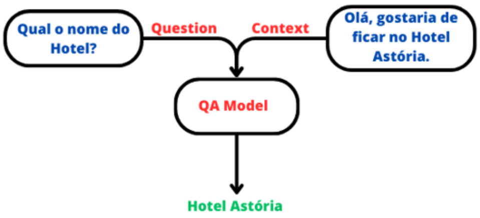


Figure 2: An Example of How a QA Model Works.

Three QA models were used: BERT-base, BERT-large, and T5. The first two are extractive models, meaning they locate answers within the provided context. In contrast, T5 is a generative model that generates its own answers. Post-Processing methods such as Levenshtein Distance (Lev) and Semantic Textual Similarity (STS) were also applied to the QA models. This step is essential because QA models do not always return answers that exactly match the original slot; sometimes there are slight variations that would be considered incorrect without Post-Processing. Some QA models also incorporate Intent Detection. Although this wasn't a primary focus of our work, we did compare results when considering all slots versus only those associated with the annotated intent. In essence, slots are linked to specific intents. When the models used have access to not only the user's sentence but also the corresponding intent, they utilize a set of questions tailored to that specific intent to fill a reduced number of slots. Conversely, models without the corresponding intent must run through a set of 30 questions, each corresponding to one of the 30 available slots for any given user sentence.

In the evaluation of the QA models, two specific metrics were used: Joint Goal Accuracy (JGA) and Slot F1. JGA serves as a comprehensive measure for assessing the system's ability to fully understand the user's overall intent. Despite its binary nature, which provides an "all-or-nothing" assessment, JGA remains essential for evaluating the effectiveness of the system in grasping the user's objectives. In contrast, the Slot F1 metric examines the system's capability to accurately identify and populate specific slots. The calculation of Slot F1 is based on both precision (the percentage of slots that the system accurately filled) and recall (the percentage of all correct slots that the system managed to identify and populate). Table 2 displays the optimal values achieved from each modification made to these three models, categorized by service.

The models yielding the best results were: BERT-base and BERT-large. The choice between these two models, as well as the specific type of Post-Processing method to use, will depend on the particular service the user aims to accomplish. As for Post-Processing techniques, our analysis did not indicate that any single method outperformed the other. Additionally, incorporating Intent Detection consistently led to significant improvements in performance across all models, highlighting its essential contribution to the system's overall performance.

| Service | QA Models with Intent Detection | JGA | F1 |
|---|---|---|---|
| Attraction | BERT-base + STS | 0.35 | 0.54 |
| Hotel | BERT-base + Lev | 0.32 | 0.50 |
| Restaurant | BERT-large + STS | 0.30 | 0.50 |
| Taxi | BERT-base + Lev | 0.47 | 0.52 |
| Train | BERT-large + Lev | 0.57 | 0.76 |

Table 2: Optimal Results Achieved by QA Models for Each Service, Incorporating Intent Detection and Post-Processing Techniques.

# 5 Conclusion

In the rapidly evolving field of dialogue systems, DST and QA techniques are crucial for dialogue systems performance. This study accomplishes its objectives through two major contributions. First, we introduce a pioneering task-oriented dataset for Portuguese. This innovation is anticipated to drive the advancement of dialogue systems in the language. A significant limitation of our approach is that the slots were frequently populated with information provided by the system during its conversation with the user. In contrast, our models could never populate these slots, as they only considered the user's sentences. The research showed that it's possible to adapt models originally designed for QA tasks to slot-filling tasks, providing a dependable solution when training data is lacking. Second, our experimental results demonstrate the significant impact of integrating Intent Detection and Post-Processing techniques. The use of Intent Detection reduces the question set from 30 to only those pertinent to the detected intent. Post-Processing techniques, on the other hand, improve the accuracy of slot-filling performed by the models. BERT models yielded the best outcomes when paired with Intent Detection. The choice between these two models, in conjunction with the most effective Post-Processing technique, will ultimately depend on the specific task that the user aims to fulfil.

Future work may include expanding the MultiWOZpt dataset. We also plan to evaluate new QA models tailored for Portuguese. Additionally, we aim to explore different Post-Processing methods and develop an Intent Detection classifier using our dataset.

**Acknowledgements**: This work is funded by the project POWER (grant number POCI-01-0247-FEDER-070365), co-financed by the European Regional Development Fund (FEDER), through Portugal 2020 (PT2020), and by the Competitiveness and Internationalization Operational Programme (COMPETE 2020).

# References

[1] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Conference on Empirical Methods in Natural Language Processing*, 2018. URL `https://api.semanticscholar.org/CorpusID:52897360`.

[2] Guan-Lin Chao and Ian Lane. Bert-dst: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. *arXiv preprint arXiv:1907.03040*, 2019.

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

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

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

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

[7] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL `https://aclanthology.org/D16-1264`.

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

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