Proposta recebida em Abril 2024 e aceite para publicação em Outubro 2024.

MultiWOZ-PT: Um Conjunto de Diálogos Orientados a Tarefas em Português

MultiWOZ-PT: A Task-oriented Dialogue Dataset in Portuguese

Patrícia Ferreira
CISUC, LASI
DEI, Universidade de Coimbra

Francisco Pais
CISUC, LASI
DEI, Universidade de Coimbra

Catarina Silva
CISUC, LASI
DEI, Universidade de Coimbra

Ana Alves
CISUC, LASI
Instituto Superior de Engenharia de Coimbra

Hugo Gonçalo Oliveira
CISUC, LASI
DEI, Universidade de Coimbra

Resumo

Apesar do amplo uso da língua portuguesa, corpos de diálogos em português, disponíveis publicamente e com anotações, são escassos. Isto torna ainda mais desafiante o desenvolvimento de sistemas de diálogo eficazes que comuniquem nesta língua. Assim, apresentamos o MultiWOZ-PT, um novo conjunto de diálogos orientados a tarefas, resultado da tradução manual de uma parte do conjunto MultiWOZ para a variedade europeia do português, e da adaptação da sua base de dados. Fornecemos diretrizes abrangentes sobre o processo de criação do MultiWOZ-PT e, para demonstrar a sua utilidade prática, realizamos experiências em dois cenários orientados a tarefas: Reconhecimento de Intenções e Monitorização do Estado do Diálogo, ambos úteis para sistemas de diálogo. Os resultados obtidos ilustram a utilidade do conjunto de dados e o seu potencial para treinar e avaliar modelos de compreensão de linguagem natural e gestão de diálogo para o português. Assim sendo, o MultiWOZ-PT constitui uma contribuição significativa para o processamento computacional dessa língua, incentivando mais pesquisas e desenvolvimento de trabalho nas áreas alvo.

Palavras chave

conjunto de diálogos orientados a tarefas; tradução; multiWOZ; monitorização do estado do diálogo; reconhecimento de intenções; preenchimento de slots

Abstract

Despite the language widespread usage, publicly available and annotated Portuguese dialogue corpora are scarce. This poses a significant challenge in the development of effective dialogue systems that communicate in Portuguese. Having this in mind, we present MultiWOZ-PT, a new task-oriented dialogue dataset that results from the manual translation of dialogues in the MultiWOZ dataset to the European variety of Portuguese, as well as an adaptation of its database. We provide comprehensive guidelines and insights into the process of creating MultiWOZ-PT and, to demonstrate its practical utility, we conducted experiments in two task-oriented scenarios: Intent Recognition and Dialog State Tracking, both useful for dialogue systems. Reported results illustrate the dataset's effectiveness and its potential for training and evaluating language understanding and dialogue management models for Portuguese. Therefore, MultiWOZ-PT constitutes a significant contribution to the computational processing of this language, fostering further research and development.

Keywords

task-oriented dialogue dataset; translation; multiWOZ; dialogue state tracking; intent recognition; slot filling

1. Introdução

A tentativa de capacitar máquinas da possibilidade de se envolverem em conversas com humanos é um dos objetivos fundamentais da Inteligência Artificial e, mais especificamente, do Processamento de Linguagem Natural (PLN). Podemos dizer que *chatbots* e assistentes virtuais são a materialização deste objetivo. Especificamente, os sistemas de diálogo orientados à tarefa são projetados para ajudar os utilizadores a realizar tarefas bem definidas em diversos domínios, tais como fazer reservas em restaurantes, consultar informações meteorológicas e agendar voos, o que os torna altamente úteis para aplicações do mundo real (Zhang et al., 2020).

Corpos de dados anotados de alta qualidade

são cruciais para o treino e avaliação de sistemas de diálogo. Desenvolver sistemas que se comunicam na língua nativa dos utilizadores é essencial para garantir acessibilidade e usabilidade, permitindo que mais pessoas interajam eficazmente com esses sistemas. Isso leva-nos a uma problema central: a escassez de corpos de dados de diálogo orientados a tarefas (CDDOTs) em idiomas que não o inglês. Quando comparados ao inglês, todos os idiomas são de recursos limitados (Wu & Dredze, 2020). Isto reflete-se na falta de recursos linguísticos para esses idiomas, e representa um desafio significativo ao limitar o desenvolvimento de sistemas orientados a dados, começando por aqueles que se baseiam em aprendizagem supervisionada.

Considerando a falta de um CDDOT disponível publicamente para o português europeu, adaptámos manualmente (parte) do Multi-Domain Wizard-of-Oz (MultiWOZ) (Budzianowski et al., 2018) para este idioma, o que resultou no MultiWOZ-PT. Este processo envolveu a adaptação da base de dados e a tradução do texto na porção de teste do MultiWOZ, que conta com 1.000 diálogos: um ficheiro com 512 e outro com 488. A base de dados original contém informações sobre serviços como hotéis, restaurantes, atrações, entre outros e foi adaptada para alinhar esses serviços, originalmente projetados para a cidade de Cambridge, Reino Unido, com o contexto de Coimbra, uma cidade em Portugal, também conhecida pela sua universidade. Para garantir consistência e qualidade dos dados traduzidos, estabelecemos algumas diretrizes, também aqui apresentadas.

O artigo demonstra ainda a utilidade prática do novo conjunto de dados. Isso é feito através da sua aplicação a duas tarefas, comuns na análise de diálogo e úteis para o desenvolvimento de sistemas de diálogo orientados a tarefas, nomeadamente: Reconhecimento de Intenções e Monitorização do Estado do Diálogo (em inglês, *Dialogue State Tracking*, DST).

Os resultados confirmam o potencial deste conjunto de dados para treinar e avaliar as tarefas-alvo em português, algo que, sem o MultiWOZ-PT, seria mais desafiante. Isso acontece porque modelos para as tarefas anteriores em português muitas vezes fazem parte de soluções proprietárias e/ou são treinados em dados proprietários, e portanto não estão disponíveis gratuitamente, nem para todos.

O restante artigo está organizado da seguinte forma. Na Secção 2, apresentamos o trabalho relacionado com foco em abordagens para criar corpos de dados anotados em português. Na Secção 3, apresentamos as diretrizes genéricas, descrevemos o conjunto de dados original MultiWOZ e a criação do MultiWOZ-PT, incluindo a sua tradução, a adaptação da base de dados e a correção de erros encontrados nas anotações originais. Na Secção 4, apresentamos e discutimos as experiências realizadas com o novo conjunto de dados, originalmente publicadas na 16ª *International Conference on Computational Processing of Portuguese* (PROPOR 2024) (Pais et al., 2024). Finalmente, na Secção 5, resumimos as principais conclusões e discutimos possíveis trabalhos futuros.

2. Trabalho relacionado

Diferentes abordagens têm sido adotadas para criar corpos de dados anotados em português. Uma passa por construí-los do zero no idioma desejado, envolvendo a recolha de documentos e sua anotação manual. Destacam-se: as coleções do HAREM (Mota & Santos, 2008; Freitas et al., 2010), onde documentos foram recolhidos de diferentes fontes e entidades mencionadas foram manualmente anotadas por especialistas na língua; ou a coleção da tarefa ASSIN (Fonseca et al., 2016), onde frases foram recolhidas e agrupadas em pares, e as suas classes de similaridade semântica e implicação foram adicionadas, com base em opiniões humanas.

A auto-anotação (em inglês, *self-labelling*) está relacionada com a anterior, mas não requer uma intervenção humana direta. É comumente adotada para corpos de dados de classificação de texto e pode depender da presença de *hashtags* específicas, por exemplo, para a classificação de ironia (da Silva, 2018); ou na fonte dos documentos, por exemplo, para a classificação de humor (Gonçalo Oliveira et al., 2020).

Uma estratégia diferente, comum para idiomas com recursos limitados, baseia-se na tradução de um conjunto de dados anotado já existente para um idioma, frequentemente inglês, para outro idioma, aproveitando as anotações existentes. Isso economiza o esforço de anotação, que leva tempo e normalmente requer especialistas ou acesso a um grande público. A tradução pode ser realizada por humanos ou com a assistência de ferramentas de tradução automática. Para o português, exemplos desta abordagem incluem a tradução de referências conhecidas para avaliar a similaridade semântica e analogia (Querido et al., 2017); ou a tradução do conjunto de dados SQuAD para resposta a perguntas (Carvalho et al., 2021).

Apesar dos exemplos anteriores, a disponibi-

lidade de corpos de dados de diálogo anotados em português é muito limitada. Por exemplo, não há um conjunto como o amplamente utilizado MultiWOZ (Budzianowski et al., 2018) ou o Schema-Guided Dialogue (SGD)(Rastogi et al., 2020), que têm conversas humanas orientadas a tarefas em inglês. Esse tipo de conjunto de dados (ou seja, CDDOT) permite aos utilizadores treinar e avaliar sistemas computacionais para uma ampla gama de tarefas no domínio do diálogo, como a geração de respostas (Gu et al., 2021), o reconhecimento de intenções (Ouyang et al., 2022; Lamanov et al., 2022), o reconhecimento de atos de diálogo (Ribeiro et al., 2019) ou a Monitorização do Estado do Diálogo (DST) (Gao et al., 2019; Siddique et al., 2021).

O MultiWOZ cobre uma diversidade de domínios e cenários, com diálogos abrangendo vários serviços, incluindo hotéis, restaurantes, atrações turísticas e transportes na cidade de Cambridge, Reino Unido. Os *slots*, que permitem armazenar informações específicas fornecidas pelos utilizadores durante a interação, estão definidos para cada serviço. O SGD referido acima também cobre diferentes domínios, cada um com esquemas e *slots* específicos. No entanto, ao contrário do MultiWOZ, cada diálogo é restrito a um único domínio.

Além do clássico MultiWOZ, este conjunto de dados tem uma versão deslexicalizada (Nekvinda & Dušek, 2021), frequentemente usada para avaliar o desempenho de sistemas de diálogo sem depender de informações específicas da base de dados. Os dados manualmente anotados capturam nuances, informações específicas e contexto detalhado, enquanto dados deslexicalizados simplificam em demasia a linguagem, resultando na perda de informação. Essa riqueza do contexto e a fidelidade ao uso da linguagem natural foi a principal razão para escolhermos dados manualmente anotados ao invés de dados deslexicalizados.

Devemos ainda referir o GlobalWOZ (Ding et al., 2021), que resulta da tradução automática do MultiWOZ para um conjunto de 20 línguas, incluindo o português. No entanto, embora a tradução automática acelere efetivamente o processo de criação, ela nem sempre será a melhor opção, como aliás mostraremos mais à frente (secção 3.3.1). Ou seja, a intervenção humana é ainda necessária para criar um conjunto de dados de qualidade, fiel a todas as nuances do diálogo e do idioma-alvo.

3. Criação do conjunto de dados

Nesta secção, descrevemos a criação do MultiWOZ-PT, um conjunto de diálogos orientados a tarefas, escritos em português. Começamos com as diretrizes seguidas para adaptar um conjunto de dados como o MultiWOZ, antes de apresentar o conjunto de dados original e o resultante, e fornecer detalhes sobre o processo de tradução e adaptação que resultou no novo conjunto de dados.

Este processo levou cerca de três meses e foi realizado por dois estudantes de Engenharia Informática (um de mestrado e um de doutoramento), autores deste artigo, falantes nativos de português europeu e com experiência em PLN, e atualmente a trabalhar em análise de diálogo. Durante este processo, reuniões com três membros seniores da equipa foram frequentes e utilizadas para rever amostras das traduções, discutir questões emergentes, e decidir como lidar com questões semelhantes no futuro.

3.1. Diretrizes gerais

Criar um novo conjunto de dados adaptado de uma fonte original é uma tarefa complexa que requer atenção meticulosa aos detalhes. Antes de iniciar o processo de tradução e adaptação, é fundamental compreender a fonte original e identificar de onde as informações serão extraídas, como bases de dados, páginas web, documentos ou outras referências relevantes.

Este trabalho teve como principal objetivo transformar um conjunto de dados em inglês, com características específicas, num conjunto de dados em português adaptado às especificidades do novo contexto linguístico, com base em fontes apropriadas e confiáveis. Para alcançar esse objetivo, traduções e adaptações precisas foram realizadas, levando em consideração as nuances da língua e da cultura.

Os tradutores humanos devem ser fluentes no idioma-alvo, pois um forte conhecimento linguístico é fundamental para manter a fidelidade e coesão no novo conjunto de dados. Além disso, entender o contexto cultural e regional-alvo é necessário para adaptar expressões, referências culturais e peculiaridades locais conforme apropriado.

Após a tradução e adaptação, deve ser realizada uma revisão abrangente das informações resultantes para garantir precisão e qualidade. Também é essencial documentar o processo de tradução e adaptação, registando decisões, recursos de referência e procedimentos específicos para ajudar a rastrear escolhas e garantir transparência no processo.

Estas diretrizes são relevantes para uma variedade de projetos de tradução e adaptação de dados, e foram aplicadas neste trabalho. Acreditamos que elas oferecem uma base sólida para garantir resultados de qualidade, independentemente da origem dos dados ou do idioma alvo.

3.2. MultiWOZ

O ponto de partida para a criação do MultiWOZ-PT foi o MultiWOZ 2.2 (Zang et al., 2020), um conhecido CDDOT. Os diálogos do MultiWOZ abrangem múltiplas interações alternadas entre dois participantes humanos: um deles assume o papel de utilizador, que tem uma tarefa a cumprir; o outro atua como um sistema, e tem como objetivo responder prontamente às solicitações do utilizador, auxiliando assim na conclusão da tarefa. É relevante notar que, no MultiWOZ, os utilizadores acreditam que estão a interagir com uma máquina (WoZ), quando, na verdade, estão a interagir com um operador humano.

A versão 2.2 do MultiWOZ foi selecionada com base na consistência e compatibilidade com modelos de dados e linguagem previamente estabelecidos, pois segue padrões de anotação para diferentes componentes dos diálogos como as intenções do utilizador e slots. Além disso, a disponibilidade de documentação e recursos facilita o desenvolvimento de pesquisas com base neste conjunto de dados. O MultiWOZ 2.2 é dividido em três partes: desenvolvimento (Desenv), treino e teste. A Tabela 1 fornece uma visão geral da distribuição do MultiWOZ 2.2 em relação à porção, número de diálogos, número de interações e média do número de interações por diálogo.

Os diálogos no MultiWOZ 2.2 abrangem um total de oito domínios: restaurante, atração, hotel, táxi, comboio, autocarro, hospital e polícia. Cada um representa uma área temática específica que molda a conversa. Outras anotações relevantes incluem intenções e *slots*. As intenções representam o objetivo subjacente por trás de uma interação do utilizador numa conversa, e os *slots* representam informações específicas que o sistema precisa de extrair das interações do utilizador para manter o contexto da conversa. O MultiWOZ original inclui 61 tipos de *slots* e tem uma estrutura organizada para as informações que podem ser solicitadas e fornecidas durante os diálogos. A Tabela 2 detalha as intenções e os *slots* no MultiWOZ. Foram utilizados os nomes originais dos *slots* e intenções, em inglês, porque não vimos necessidade para os traduzir. Assim também fica mais fácil a correspondência com as anotações do MultiWOZ original.

Além dos diálogos, inclui uma base de dados com informações sobre serviços na cidade de Cambridge, Reino Unido, relacionados aos domínios cobertos. A maioria das respostas do sistema às solicitações dos utilizadores resulta da aplicação das restrições dos utilizadores (i.e., valores dos *slots*) às informações nesta base dados.

Os *slots* servem para manter, de forma estruturada, a informação acerca do contexto. Eles podem ser categóricos ou não categóricos. Os valores dos *slots* categóricos são limitados a um conjunto de valores pré-definidos. Por exemplo, no serviço de hotel, os valores válidos para *type* e *pricerange* são respetivamente: *guesthouse* ou *hotel* e *expensive*, *cheap* ou *moderate*. Por outro lado, os *slots* não categóricos aceitam valores contínuos ou texto livre. Por exemplo, o *slot address* refere-se à morada do serviço e pode conter tanto valores numéricos quanto informações textuais.

3.3. MultiWOZ-PT

O MultiWOZ-PT resulta da tradução e adaptação do conjunto de dados original do MultiWOZ para português. Devido às limitações de tempo inerentes à metodologia que adotamos, que depende fortemente do trabalho manual, tivemos de começar por traduzir apenas parte dos diálogos originais. Nesse sentido, optamos por uma partição já existente, que poderia ter sido a de desenvolvimento ou de teste, cada uma com 1.000 diálogos. Acabou por não haver uma razão especial para termos optado pela porção de teste, mas o objetivo seria sempre, quando necessário, separar essa partição em conjuntos de treino e teste. Neste caso, a própria partição de teste é constituída por dois ficheiros que podem ser usados para o efeito[1]. Além disso, em futuros trabalhos, pretendemos expandir o MultiWOZ-PT, começando pela tradução e adaptação dos dados de desenvolvimento do MultiWOZ original.

A porção de teste do MultiWOZ 2.2 possui então 1.000 diálogos e é originalmente dividida em dois ficheiros, um com 512 diálogos e outro com 488. Na Secção 4, utilizamos esses ficheiros para treino e teste, respetivamente.

Esses 1.000 diálogos não cobrem todos os domínios, nem todos os tipos de *slot* do MultiWOZ. Eles limitam-se a cinco (restaurante, atração, hotel, táxi, comboio) dos oito domínios

[1]Admitimos que a escolha da partição de teste pode afetar a avaliação comparativa com estudos multilíngues que mantiveram o conjunto de teste original para a validação final.

2-6	#Diálogos	#Interações			#Interações
		Total	Utilizador	Sistema	/Diálogo
Treino	8.436	113.552	56.776	56.776	13,46
Desenv	1.000	14.748	7.374	7.374	14,74
Teste	1.000	14.744	7.372	7.372	14,75

Tabela 1: Distribuição do conjunto de dados MultiWOZ 2.2 nas porções de desenvolvimento (Desenv), treino e teste.

Domínio	*Slots* Categóricos	*Slots* Não Categóricos	Intenções
Restaurante	*pricerange*, *area*, *bookday*, *bookpeople*	*food*, *name*, *booktime*, *address*, *phone*, *postcode*, *ref*	*find*, *book*
Atração	*area*, *type*	*name*, *address*, *entrancefee*, *openhours*, *phone*, *postcode*	*find*
Hotel	*pricerange*, *parking*, *internet*, *stars*, *area*, *type*, *bookpeople*, *bookday*, *bookstay*,	*name*, *address*, *phone*, *postcode*, *ref*	*find*, *book*
Taxi	–	*destination*, *departure*, *arriveby*, *leaveat*, *phone*, *type*	*book*
Comboio	*destination*, *departure*, *day*, *bookpeople*	*arriveby*, *leaveat*, *trainid*, *ref*, *price*, *duration*	*find*, *book*
Autocarro	*day*	*departure*, *destination*, *leaveat*	*find*
Hospital	–	*department*, *address*, *phone*, postcode	*find*
Polícia	–	*name*, *address*, *phone*, *postcode*	*find*

Tabela 2: Mapeamento de *slots* e intenções para os domínios do MultiWOZ.

originais e a 30 tipos de *slot*. Especificamente, das 1928 interações, 399 estão relacionadas a atrações, 396 a hotéis, 445 a restaurantes, 198 a táxis e 490 a comboios. Ao traduzir, mantivemos a consistência semântica para preservar as intenções, domínios e valores dos *slots* originais. É de notar também que os números de referência originais para reservas de serviços foram preservados. No resto da secção, aprofundamos os detalhes da nossa metodologia, focando nos dois principais passos.

3.3.1. Tradução

Todas as interações na porção de teste do MultiWOZ foram traduzidas manualmente do idioma de origem, o inglês, para o português. A Tabela 3 complementa a ilustração do MultiWOZ-PT, desta vez com um diálogo completo do MultiWOZ original (em inglês) e a sua tradução para o português, no MultiWOZ-PT.

Vale ressaltar que uma tradução direta, palavra por palavra, da maioria das interações em inglês não seria viável, pois resultaria em traduções pouco naturais.

Portanto, os tradutores humanos tiveram de compreender o significado fundamental de cada interação nos diálogos originais e transmiti-lo com precisão em português, seguindo as diretrizes do mais recente Acordo Ortográfico (Silva, 2012).

Expressões idiomáticas foram adaptadas para garantir naturalidade e compreensão para os falantes de português europeu. Por exemplo, a interação "Let's grab a bite at this joint" foi traduzida para "Vamos petiscar neste local" e a interação "I'm heading downtown" foi traduzida para "Estou a caminho da baixa da cidade", refletindo melhor o contexto cultural e linguístico de Portugal. Isso garantiu que os diálogos traduzidos não só mantivessem o seu significado, mas refletissem também as nuances e peculiaridades culturais, regionais e contextuais da língua portuguesa, uma tarefa improvável de ser alcançada por meio de tradução automática. Esta abordagem centrada no ser humano é a chave para criar um conjunto de dados de alta qualidade para tarefas de PLN relacionadas a diálogos em português.

Isso é ainda mais evidente ao comparar praticamente qualquer diálogo na tradução para o português no GlobalWOZ (Ding et al., 2021) com o diálogo correspondente no MultiWOZ-PT. A Tabela 4 ilustra esse facto com algumas in-

terações retiradas de ambos os conjuntos de dados.

De forma a quantificar os problemas no GlobalWOZ, foi selecionada uma amostra aleatória de 30 diálogos deste conjunto, totalizando 402 interações, nas quais foram identificadas diversas inconsistências categorizadas em nove tipos diferentes. A Tabela 5 apresenta um resumo dessas inconsistências, indicando quantas vezes cada tipo específico foi observado nas interações de cada diálogo analisado, juntamente com um exemplo representativo de cada tipo. As inconsistências foram agrupadas em categorias como omissão de palavras e/ou frases que estão na interação original, inconsistência gramatical (inclui falta de artigos definidos, preposições incorretas, entre outros), tradução errada, inconsistência de género, ordem incorreta de palavras, falta de pontuação nas interações, contexto de Cambridge em vez de Lisboa e repetição de palavras. A média de inconsistências por interação nesta amostra é de 1,37.

Nestes 30 diálogos, verificamos que 86 interações apresentam claramente a variedade brasileira, enquanto que nós estamos a traduzir para o português europeu. Um exemplo disso é a interação "estou precisando de um trem saindo na terça-feira e chegando às 20:39". Observámos também que, em várias interações no GlobalWOZ, a interação do sistema não responde ao pedido do utilizador. Por exemplo, o utilizador pergunta "há alguma piscina a leste?" e o sistema responde "sim , existem dois Bares e discotecas .", o que indica que, em alguns casos, a base de dados de Cambridge não foi adaptada com os mesmos tipos de serviço. Este é um aspeto muito importante que foi sempre considerado na adaptação da base de dados do MultiWOZ-PT.

3.3.2. Adaptação da Base de Dados

Como mencionado anteriormente, as respostas do sistema no MultiWOZ são baseadas numa base de dados com serviços de Cambridge, Reino Unido. No entanto, ter tantos diálogos em português centrados nos serviços em Cambridge, incluindo localizações e contactos, não seria muito natural. Portanto, além de traduzirmos as interações, personalizamos os diálogos para o contexto específico da cidade portuguesa de Coimbra.

A escolha de Coimbra deve-se ao facto de ser uma cidade em Portugal, com algumas semelhanças a Cambridge, principalmente devido à sua universidade.

Não foram introduzidos tipos de *slots* além dos já existentes no conjunto de dados original. No entanto, os valores dos *slots* não categóricos foram adaptados para refletir os serviços em Coimbra. Ao adaptar restaurantes, atrações e hotéis, alteramos sistematicamente os campos essenciais, como *address*, *name*, *phone* e *postcode*, garantindo a sua consistência e relevância no novo local. As informações anteriores foram obtidas maioritariamente de plataformas conhecidas na rede, como Tripadvisor[2], principalmente para restaurantes, Booking[3], principalmente para hotéis, e também dos sites oficiais dos serviços como os sites dos próprios restaurantes, atrações ou hotéis para obter informações adicionais.

Os campos restantes foram adaptados conforme necessário para serviços específicos. Para hotéis, os campos adaptados, quando necessário, incluíram *area*, *internet*, *parking*, *pricerange*, *stars* e *type*. No caso de restaurantes, os campos adaptados, quando necessário, abrangeram *area*, *food* e *pricerange*. Para atrações, os campos adaptados incluíram *area*, *entrancefee*, *openhours*, *pricerange* e *type*. No que diz respeito aos comboios, as informações foram obtidas no sítio da CP[4], que é a plataforma da companhia ferroviária nacional de Portugal. Ele fornece informações sobre horários de comboio, rotas, preços e serviços. O nosso objetivo foi escolher comboios que tivessem como destino ou partida a cidade de Coimbra, e com durações semelhantes às da base de dados original, sempre que possível.

Para ilustrar o processo de adaptação da base de dados do MultiWOZ para o contexto de Coimbra, a Tabela 6 apresenta exemplos específicos de entradas ajustadas para dois tipos de serviços: hotel e restaurante. Esta tabela mostra uma entrada representativa para cada serviço originalmente localizado em Cambridge e como foi adaptada para a nova localização em Coimbra, mantendo a estrutura e a informação relevantes para cada categoria de serviço.

Além disso, o conjunto de dados original continha um número significativo de restaurantes indianos em Cambridge, que não são tão comuns em Coimbra. Para fazer as adaptações necessárias, seguimos duas abordagens distintas. Primeiro, adicionamos um sufixo aos nomes dos restaurantes indianos em Coimbra na nossa base de dados de restaurantes (por exemplo, *Italian Indian Palace* tornou-se *Italian Indian Palace 2* ou *Chai...*

[2] https://www.tripadvisor.pt/
[3] https://booking.com/
[4] https://www.cp.pt/

Interlocutor	MultiWOZ 2.2	MultiWOZ-PT
UTILIZADOR	I need info on a train that would be departing from Peterborough.	Preciso de informações sobre um comboio que parta da Figueira da Foz.
SISTEMA	What day and time?	A que dia e hora?
UTILIZADOR	I would like to leave on Sunday and arrive in Cambridge by 15:15.	Gostaria de partir no domingo e chegar a Coimbra pelas 15:15.
SISTEMA	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?
UTILIZADOR	That'd be perfect, I need three tickets on sunday.	Isso seria perfeito, preciso de três bilhetes para domingo.
SISTEMA	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?
UTILIZADOR	I'm also looking for information on a place called the man on the moon.	Também estou à procura de informações sobre um lugar chamado Salão Brazil.
SISTEMA	The man on the moon is a concerthall in the centre. Would you like to know anything else?	O Salão Brazil é uma sala de concertos no centro. Gostaria de saber mais alguma coisa?
UTILIZADOR	Yes can I get the address to that please?	Sim, pode dar-me o endereço, por favor?
SISTEMA	Yes, the address is 2 norfolk street.	Sim, o endereço é Largo do Poço, nº3, 1º Andar.
UTILIZADOR	Can you tell me what the area of town is? Can I get the phone number too please?	Pode dizer-me qual é a zona da cidade? Pode também dar-me o número de telefone, por favor?
SISTEMA	It's located in the centre 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?
UTILIZADOR	That will be it for today! Thank you so much!	É tudo por hoje! Muito obrigado!
SISTEMA	You're very welcome! Have a great day!	Não tem de quê! Tenha um ótimo dia!

Tabela 3: ID do diálogo original PMUL1241 do MultiWOZ e a sua tradução no MultiWOZ-PT.

Conjunto de dados	Interação
MultiWOZ	Please find a moderate price range Turkish restaurant.
GlobalWOZ	encontre um restaurante de faixa de preço preço moderado Europeia .
MultiWOZ-PT	Por favor, procure um restaurante turco de preço moderado.
MultiWOZ	I want to book a table for 3 people at 13:00 on wednesday will you do that ?
GlobalWOZ	eu quero reservar uma mesa para pessoas Sete em 19:00 em quarta-feira .
MultiWOZ-PT	Quero reservar uma mesa para 3 pessoas às 13:00 de quarta-feira, podem fazê-lo?
MultiWOZ	Would you help me book a train today?
GlobalWOZ	você me ajudaria a reservar um trem hoje ?
MultiWOZ-PT	Poderia ajudar-me a reservar um comboio hoje?
MultiWOZ	I would be happy to. Where are you traveling?
GlobalWOZ	Eu ficaria feliz em .
MultiWOZ-PT	Teria todo o gosto. Para onde viaja?

Tabela 4: Interações no MultiWOZ original (diálogos PMUL3731 e PMUL1623), GlobalWOZ e MultiWOZ-PT.

Inconsistência	Contagem	Contagem (%)	Exemplo da interação com inconsistência e da interação original do MultiWOZ
Omissão de parte da frase	216	46.65	train PL020 atenderia aos seus critérios. Train TR4803 would meet your criteria. Can I book something for you?
Inconsistência gramatical	132	28.51	posso fazer uma reserva para nove no 17:30 nesse quarta-feira? Can I book a reservation for 7 at 13:00 this Friday?
Tradução errada	37	7.99	por favor, encontre um lugar para ir na leste e deve ser uma colagem. please find me a place to go in the centre and it should be a collage.
Inconsistência de género	33	7.13	claro, há dois na Oeste. Sure there are two in the centre of town. I prefer Anatólia located at 30 Bridge Street City Centre.
Ordem incorreta	20	4.32	sim, para pessoas Sete. Yes for 6 people. 2 nights starting from saturday.
Falta de pontuação	9	1.94	Preciso reservar uma mesa para o mesmo grupo de pessoas em 20:00 no terça em um restaurante Europeia. I need to book a table for the same group of people at 12:30 on the same day at a mexican restaurant.
Contexto de Cambridge	8	1.73	estou procurando um lugar em Cambridge. I am looking for a place in Cambridge. It doesn't need to include internet and should be a hotel
Repetição de palavras	8	1.73	estarei entrando na central de central e quero chegar por volta de central. I'll be coming into Cambridge from Bishops Stortford.

Tabela 5: Exemplos e contagem de inconsistências encontradas numa amostra de 30 diálogos do GlobalWOZ e a respetiva interação no MultiWOZ original.

Serviço	MultiWOZ 2.2	MultiWOZ-PT
Hotel	name: ashley hotel address: 74 chesterton road area: north internet: yes parking: yes phone: 01223350059 postcode: cb41er pricerange: moderate stars: 2 type: hotel	name: Hotel Mondego address: Largo das Ameias 3-4 area: north internet: yes parking: no phone: 239496239 postcode: 3000-024 pricerange: moderate stars: 2 type: hotel
Restaurante	name: ask restaurant address: 12 Bridge Street City Centre area: centre food: italian phone: 01223364917 postcode: cb21uf pricerange: cheap type: restaurant	name: La Divina Pizza Bar address: Rua Carlos Seixas 267 area: centre food: italian phone: 239093914 postcode: 3030-177 pricerange: cheap type: restaurant

Tabela 6: Exemplos de serviços da base de dados do MultiWOZ para Cambridge e as suas adaptações no MultiWOZ-PT para Coimbra.

tornou-se *Chai2...*). Uma segunda opção envolveu a substituição de referências a certos restaurantes indianos por estabelecimentos de outro tipo mais prevalentes em Coimbra (por exemplo, o restaurante indiano *pipasha restaurant* foi substituído por um restaurante português "Casas do Bragal").

3.3.3. Correção dos procedimentos

Algumas palavras com erros ortográficos em inglês foram identificadas no conjunto de dados original. Por exemplo, a palavra *avaliable* na interação *There are several guesthouses avaliable. (...)* (diálogo MUL2155) ou a palavra *adress* na interação *get me the adress please* (diálogo PMUL3282). Apesar de ser importante que os sistemas de diálogo sejam tolerantes a este tipo de erros, eles não foram replicados na tradução para o português e as palavras foram corretamente traduzidas.

Foi ainda identificado um erro sistemático no conjunto de dados original do MultiWOZ: um total de 975 interações de utilizador foram rotuladas com múltiplas intenções, embora, na maioria dos casos, isso não devesse ocorrer. Por exemplo, a interação *I need a train from Norwich to Cambridge on Friday, please. I need to get there at about 10:30.* (diálogo PMUL4034) foi rotulada com as intenções *find-train* e *find-hotel*, embora a última fosse claramente incorreta. Essas inconsistências foram prontamente corrigidas, o que explica a disparidade no número de intenções entre o conjunto de dados original e o conjunto de dados em português, na Tabela 7.

4. Experimentação

Nesta secção, demonstramos a utilização do MultiWOZ-PT em duas tarefas relacionadas com análise de diálogo, nomeadamente: Reconhecimento de Intenções e Monitorização de Estado de Diálogo (DST, em inglês). Embora reconheçamos que existem diversas outras tarefas importantes no domínio do diálogo, escolhemos estas pois o MultiWOZ oferece suporte de anotação e é frequentemente utilizado para avaliá-las (Lamanov et al., 2022; Tavares et al., 2023; Zhang et al., 2019). Com o novo conjunto de dados, ficam disponíveis e adaptadas para o contexto da língua portuguesa, tornando-se mais acessíveis à comunidade.

4.1. Reconhecimento de Intenções

O Reconhecimento de Intenções é uma tarefa comum em sistemas de diálogo que tem como objetivo identificar aquilo que o utilizador expressa querer realizar. Assim, desempenha um papel crucial em aprimorar a experiência do utilizador.

As frases do MultiWOZ-PT estão anotadas com oito intenções diferentes, apresentadas na Tabela 7. Portanto, esta experiência envolveu o treino de um modelo para classificar automaticamente essas intenções. Para esse fim, utilizamos como ponto de partida dois modelos da família BERT disponíveis, nomeadamente: BERTimbau-base (Souza et al., 2020) e Albertina-PTPT (Rodrigues et al., 2023). Eles foram utilizados através da biblioteca *transformers*, em Python, e do repositório HuggingFace[5][6]. Ambos os modelos são dedicados ao português. No entanto, o primeiro, que está disponível há mais tempo, foi pré-treinado a partir de textos em português do Brasil. O segundo, Albertina-PTPT, é baseado no DeBERTa (He et al., 2020) e foi pré-treinado a partir de textos em português europeu, a mesma variedade do MultiWOZ-PT.

Para realizar o Reconhecimento de Intenções, os modelos anteriores foram afinados (em inglês, *fine-tuned*) nas anotações de intenção do MultiWOZ-PT. Para uma divisão mais natural, nesta fase usamos o primeiro ficheiro, que contém 512 diálogos. Os seguintes hiperparâmetros foram usados para a afinação de ambos os modelos: *batch size* 32, *learning rate* de $1e^{-5}$ e duração do treino de 5 épocas. O desempenho foi medido nos 488 diálogos restantes.

Além dos modelos baseados em BERT, também implementámos um modelo de máquina de vetores de suporte (SVM) linear com vetores Frequência do Termo-Inverso da Frequência nos Documentos (Tf-idf). Utilizou-se o SVM com kernel linear, mantendo os parâmetros padrão. Na implementação do Tf-idf, consideraram-se todos os unigramas que aparecem em pelo menos dois documentos (min_df = 2), com os restantes parâmetros configurados como padrão. Ambos estão disponíveis na biblioteca *scikit-learn*[7][8].

[5]BERTimbau disponível em huggingface.co/neuralmind/bert-base-portuguese-cased

[6]Albertina disponível em huggingface.co/PORTULAN/albertina-ptpt

[7]SVM disponível em https://scikit-learn.org/stable/modules/svm.html

[8]Tf-idf disponível em https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

	MultiWOZ		MultiWOZ-PT	
Intents	#	%	#	%
find_attraction	1.232	17,10	1.065	17,10
find_hotel	1.235	17,18	1.093	17,57
book_hotel	483	6,71	432	6,93
find_restaurant	1.183	16,42	1.035	16,62
book_restaurant	580	8,06	440	7,06
find_taxi	495	6,86	416	6,68
find_train	1.562	21,69	1.379	22,17
book_train	431	5,98	366	5,87
Total	7.201	100	6.226	100

Tabela 7: Intenções no conjunto de teste para o MultiWOZ e para o MultiWOZ-PT.

A Tabela 8 reporta a precisão (P), abrangência (R) e medida F1 (F1) dos modelos, quando aplicados a esses diálogos, em média e distribuídos por intenção.

Os modelos demonstram consistentemente um desempenho forte na tarefa de Reconhecimento de Intenções, alcançando a precisão, abrangência e medida F1 superiores ou igual a 0,77. O modelo SVM linear utilizando TF-IDF tem um desempenho ligeiramente mais baixo que os restantes que têm desempenhos semelhantes. O principal objetivo desta experiência é demonstrar o MultiWOZ-PT no Reconhecimento de Intenções orientadas em diálogos orientados a tarefas, onde pode ser usado para treinar e avaliar modelos. Como se pode verificar na Tabela 7, o conjunto de dados não é equilibrado em termos de intenções, o que pode ter impacto no desempenho. Por exemplo, os modelos Albertina-PTPT e BERTimbau têm um desempenho melhor para intenções como *find_train* (F1$\approx 0,90$), seguido por *find_attraction* e *find_hotel*, que são as três intenções mais representadas. Por outro lado, o desempenho é pior para *book_hotel* e *book_train*, as menos representadas.

4.2. DST

A maioria dos sistemas de diálogo orientados a tarefas adota uma arquitetura conhecida como Estado do Diálogo (em inglês, *Dialog State*), que compreende a Monitorização do Estado do Diálogo (DST), uma Política de Diálogo e Atos de Diálogo. O DST (Williams et al., 2016) acompanha o estado atual de uma conversa preenchendo *slots* específicos. Por exemplo, supondo que pretendemos preencher os slots *food* e *area* e que o utilizador refere a frase "Estou à procura de um restaurante europeu na zona oeste de Coimbra", podemos preencher o slot *food* com o valor "europeu" e o slot *area* com o valor "oeste". Isto realiza-se para todos os turnos do diálogo.

O objetivo desta segunda experiência é realizar DST no MultiWOZ-PT, o que envolve extrair *slots* e os seus valores a partir das interações dos utilizadores. A experiência é descrita em várias subsecções, onde também explicamos diferentes etapas da abordagem adotada.

4.2.1. Resposta a Perguntas para Preenchimento de Slots

Abordamos o preenchimento de *slots* como uma tarefa de Resposta Automática a Perguntas (QA), que pode aproveitar modelos disponíveis baseados na arquitetura *transformer*, como o BERT (Devlin et al., 2019) e o T5 (Raffel et al., 2020), afinados em corpos de dados de QA. Dada uma pergunta em linguagem natural e um contexto, estes modelos extraem a sequência no contexto que melhor responde à pergunta[9]. O preenchimento de cada *slot* baseia-se depois na realização uma pergunta em linguagem natural, utilizando a interação como contexto.

Para QA utilizamos uma versão do BERTimbau afinada numa tradução do conjunto de dados SQuAD (Rajpurkar et al., 2016) para o português, um modelo disponível no portal Huggingface[10]. O SQuAD contém mais de 100.000 pares de perguntas e respostas em mais de 500 artigos da Wikipédia. Como não é específico de um domínio, ele adapta-se bem a diferentes contextos, principalmente quando as perguntas feitas são simples e diretas.

[9]Mais propriamente, o BERT identifica o início e o fim da sequência que responde à pergunta, enquanto que o T5 gera essa sequência. Como, no SQuAD, as respostas são sempre sequências que ocorrem no contexto, é expectável que a sequência gerada pelo T5 também esteja no contexto, ainda que isso não seja garantido.

[10]https://huggingface.co/pierreguillou/bert-large-cased-squad-v1.1-portuguese

Intenção	Albertina-PTPT			BERTimbau			SVM com Tf-idf		
	P	R	F1	P	R	F1	P	R	F1
find_attraction	0,76	0,90	0,83	0,81	0,90	0,85	0,71	0,89	0,79
find_hotel	0,80	0,83	0,81	0,81	0,84	0,82	0,85	0,77	0,80
book_hotel	0,72	0,76	0,74	0,78	0,75	0,76	0,89	0,68	0,77
find_restaurant	0,84	0,77	0,80	0,87	0,75	0,81	0,75	0,77	0,76
book_restaurant	0,78	0,88	0,83	0,72	0,84	0,78	0,84	0,65	0,73
find_taxi	0,95	0,74	0,83	0,80	0,82	0,81	0,96	0,77	0,85
find_train	0,92	0,89	0,90	0,91	0,88	0,89	0,84	0,92	0,88
book_train	0,83	0,63	0,72	0,79	0,75	0,77	0,71	0,71	0,71
Média macro	0,82	0,80	0,81	0,81	0,82	0,81	0,82	0,77	0,79
Média ponderada	0,83	0,83	0,83	0,83	0,83	0,83	0,81	0,80	0,80

Tabela 8: Desempenho do Reconhecimento de Intenções no MultiWOZ-PT.

Além do modelo, é necessário ter uma pergunta em linguagem natural para cada um dos 30 tipos de *slot* no MultiWOZ-PT. As perguntas foram criadas manualmente, mas isso só tem de ser feito uma vez para cada conjunto de dados/*slots*. As perguntas foram o resultado de testes preliminares com o intuito de seguir um formato semelhante ao do SQuAD, garantindo que fossem diretas e mencionando sempre o nome do slot. Apesar de envolver trabalho manual, tem apenas de ser feito uma vez, e eventualmente revisto quando os tipos de *slot* mudam, por exemplo, para novos diálogos ou para um novo conjunto de dados. A Tabela 9 mostra as perguntas usadas para os *slots* do serviço de hotel.

A Figura 1 ilustra a abordagem adotada que representa uma contribuição deste trabalho. Dada uma pergunta ("Qual o nome do Hotel?") e uma interação ("Olá, gostaria de ficar no Hotel Astória.") como entrada, o modelo extrai a resposta desejada ("Hotel Astória"), que é então usada para preencher um *slot*, neste caso, *hotel_name*.

À sequência que responde à pergunta, os modelos de QA juntam um valor entre 0 e 1, num campo "score", associado à confiança na resposta dada. Tiramos partido desse valor para ignorar respostas cuja confiança é mais baixa. Na maior parte dos casos, estas correspondem a perguntas para as quais a interação, dada como contexto, não tem uma resposta. Na Figura 1, está um caso em que se faz uma pergunta para a qual a resposta existe e, por essa razão, a confiança é alta (0,992). Por outro lado, se fosse feita uma pergunta para a qual não há resposta, por exemplo, se a pergunta "Em que dia é a reserva?" for feita ao contexto da figura, é normal obter uma resposta com baixa confiança, tal como gostaria de ficar no "Hotel Astória", com "score" 0,069.

Embora não seja a abordagem mais comum, uma vantagem é que esta se baseia em modelos de

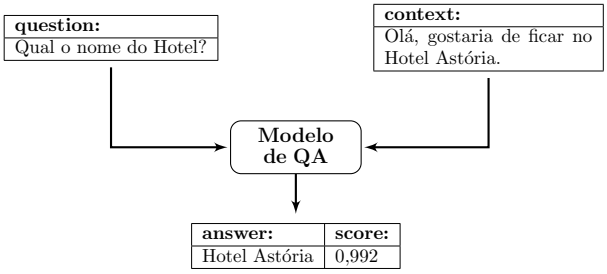

Figura 1: Exemplo do modelo de QA em funcionamento.

QA genéricos disponíveis e não requer supervisão em dados de DST. O trabalho manual necessário é mínimo, o que permite a aplicação da abordagem a cenários onde ou não há dados anotados disponíveis, ou não são suficientes para o treino. Ressalvamos que o MultiWOZ-PT possui 30 tipos diferentes de *slots* e treinar um modelo para preenchê-los exigiria mais dados do que para o Reconhecimento de Intenções.

4.2.2. Pós-processamento para Slots Categóricos

Entre as perguntas respondidas de forma imprecisa, algumas estão suficientemente próximas dos valores válidos dos *slots*. Isto acontece, por exemplo, porque os utilizadores podem usar variações como sinónimos; ou porque o modelo de QA pode extrair informações adicionais. Para mapear as respostas anteriores em valores válidos dos *slots*, é aplicado um pós-processamento aos valores dados para os *slots* categóricos.

Dois métodos diferentes foram testados para associar as respostas aos valores mais próximos para esses *slots*, nomeadamente:

- Distância de Levenshtein, baseada no número mínimo de edições de caracteres necessárias para transformar a resposta num valor válido;

Tipo de *Slot*	Pergunta
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?*

Tabela 9: Perguntas para o Serviço de Hotel.

- Similaridade Textual Semântica (STS), baseada na similaridade do cosseno entre as representações vetoriais (em inglês, *embeddings*) da resposta e cada valor válido. As representações foram obtidas a partir de um *transformer* disponível na HuggingFace, baseado na afinação do BERTimbau nos pares das coleções ASSIN[11].

A Tabela 10 ilustra o pós-processamento com exemplos onde as respostas foram mapeadas com sucesso para valores válidos de *slot*. Sem o pós-processamento, os resultados obtidos seriam considerados incorretos.

Método	*Slot* Original	*Slot* preenchido
Levenshtein	arquitetura	arquitetónico
	barco	barcos
	caro	cara
STS	meio dia	depois do meio dia
	sexta-feira	sexta-feira às 16:00
	centro	centro da cidade

Tabela 10: Exemplos de comparações corretamente avaliadas utilizando os métodos Levenshtein e STS como etapa de pós-processamento.

4.2.3. Limitação do Número de Respostas

Independentemente da pergunta, havendo um contexto, o modelo de QA extrairá sempre uma resposta. Portanto, fazer todas as 30 perguntas para cada interação do utilizador é propício a produzir ruído. Uma primeira verificação na qualidade da resposta baseia-se na pontuação que o modelo atribui à resposta (*score*), um número no intervalo de 0 a 1 que reflete a confiança. Ao definir um limite na pontuação, podemos limitar o número de respostas a serem consideradas. Um limite alto deve manter as respostas boas e descartar as de qualidade inferior, possivelmente incorretas. Um limite baixo pode ser muito restritivo e resultar em nenhuma resposta considerada. Para determinar os limites que maximizam o desempenho, testamos valores na gama de 0,49 a 0,79. Tal como na experiência anterior, isso foi feito nos primeiros 512 diálogos do MultiWOZ-PT.

Importa referir que o Reconhecimento de Intenções pode ser útil para restringir o conjunto de perguntas àquelas relacionadas com a intenção alvo. Portanto, para identificar as intenções do utilizador, usamos um dos classificadores descritos na Secção 4.1. Como o desempenho foi semelhante, escolhemos o classificador baseado no BERTimbau, que é também o modelo com menos parâmetros.

4.2.4. Avaliação

A avaliação do DST recorre geralmente a duas métricas, nomeadamente:

- A *Joint Goal Accuracy* (JGA) mede a compreensão geral do sistema em relação à intenção do utilizador. É uma métrica "tudo ou nada" que procura avaliar se o sistema compreende efetivamente todos os objetivos do utilizador.
- A *Slot* F1 avalia a capacidade do sistema de identificar com precisão os valores dos *slots* associados a cada interação

A abordagem baseada em QA foi testada na segunda parte do MultiWOZ-PT, compreendendo 488 diálogos. A Tabela 11 relata o seu desempenho ao aplicar cada um dos métodos de pós-processamento ou nenhum.

O DST é desafiador e os valores de referência no MultiWOZ são inferiores a 0,60 (Heck et al., 2020; Lee et al., 2021). Como a abordagem adotada não foi supervisionada nesta tarefa, consideramos os resultados obtidos como interes-

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

santes e como uma base que pode ser melhorada no futuro. Concluímos também que o pós-processamento é benéfico e que ambos os métodos têm uma JGA comparável, ainda que a distância de Levenshtein leve a pontuações de *Slot* F1 superiores.

Diferentes serviços apresentam desafios diferentes e ambas as métricas variam entre eles. O melhor desempenho é alcançado para os comboios. O pior depende da métrica e é para restaurante (JGA) e táxi (*Slot* F1).

5. Conclusões e Trabalho Futuro

Apresentamos o MultiWOZ-PT, um conjunto de dados de diálogo orientados a tarefas criado a partir do conjunto MultiWOZ, amplamente utilizado, que adaptámos para o contexto da língua e cultura portuguesas. Este novo conjunto de dados é uma resposta à ausência de um CDDOT de qualidade e publicamente disponível em português. Os diálogos e a base de dados do MultiWOZ-PT estão disponíveis para todos os interessados no repositório GIT[12] e no repositório OSF[13], onde também está incluído o código produzido para a realização das experiências aqui descritas.

O processo de adaptação envolveu: (i) a tradução e contextualização da parte de teste do MultiWOZ; e (ii) a adaptação da base de dados original, onde que serviços em Cambridge, Reino Unido, foram substituídos por serviços semelhantes em Coimbra, Portugal.

Além de apresentarmos o MultiWOZ-PT, realizámos duas experiências que ilustram possíveis aplicações do conjunto de dados e confirmam que ele abre portas a tarefas relacionadas com o processamento de diálogos em português. Começámos com o Reconhecimento de Intenções, um passo comum na maioria dos sistemas de diálogo, útil para diferenciar o tratamento de diferentes pedidos e consequentes respostas. Em seguida, abordámos o DST, útil para representar o contexto de um diálogo orientado a tarefas, ao procurar respostas plausíveis, considerando mais do que apenas a última interação. Ambas as experiências foram realizadas com modelos reconhecidos e confirmaram a coerência do novo conjunto de dados.

Este trabalho aborda uma lacuna crítica no processamento computacional da língua portuguesa, atuando como um catalisador para futuras pesquisas e avanços. Como um conjunto de dados publicamente disponível, o MultiWOZ-PT permite o treino e/ou a avaliação de novos modelos para as tarefas anteriores ou outras relacionadas com o processamento de diálogo na língua portuguesa. Assim, esperamos estar a contribuir para o avanço do estado da arte. Com um melhor processamento da língua portuguesa, estamos a beneficiar não apenas os falantes dessa língua, mas também a fortalecer o desenvolvimento de tecnologias mais acessíveis e eficazes.

Tendo em vista uma referência mais abrangente, o trabalho futuro inclui o aumento do tamanho do conjunto de dados. Isto exigirá a tradução de mais diálogos, ao mesmo tempo que serão feitas adaptações necessárias na base de dados. Em vez de seguir o mesmo processo, podemos explorar abordagens automáticas de adaptação, como a do GlobalWOZ (Ding et al., 2021). No entanto, ainda é nossa intenção tratar manualmente os diálogos e a base de dados resultantes, e aproveitar essa abordagem apenas para acelerar o processo. Na verdade, com os problemas observados na tradução para o português do GlobalWOZ, isso é altamente recomendável.

Como mostramos, o MultiWOZ-PT pode suportar diferentes tarefas relacionadas à análise de diálogos em português. Os resultados apresentados aqui podem ser vistos como linhas de base e têm espaço para melhorias, algo que também planeamos abordar no futuro. Por exemplo, o Reconhecimento de Intenções pode beneficiar da consideração das intenções anteriores na classificação (por exemplo, adicionando uma camada CRF aos modelos atuais). Quanto ao DST, o próximo passo seria testar abordagens supervisionadas no MultiWOZ-PT, começando com aquelas já aplicadas ao MultiWOZ (Heck et al., 2020; Lee et al., 2021).

Algumas das abordagens anteriores são baseadas na afinação de um modelo de linguagem para as respetivas tarefas. Aqui, uma experiência interessante seria analisar o impacto do treino de modelos numa quantidade maior de diálogos traduzidos automaticamente por oposição a uma quantidade menor de diálogos traduzidos manualmente. Para garantir uma avaliação justa, os modelos seriam sempre comparados em diálogos traduzidos manualmente, como os que já temos.

Por fim, também estamos interessados em analisar fluxos comuns em diálogos em português de diferentes tipos, com múltiplos propósitos. Aqui, mais do que fazer essa análise no MultiWOZ-PT, o novo conjunto de dados pode apoiar alguns de nossos estudos. Mais precisamente, modelos treinados neste conjunto de dados podem ser usados para enriquecer diálogos

[12] https://github.com/NLP-CISUC/Dialog-State-Tracking-PT

[13] https://osf.io/vxq6p/

Serviços	JGA			*Slot* F1		
	Nenhum	Lev	STS	Nenhum	Lev	STS
Atração	0,25	0,36	**0,37**	0,46	**0,53**	0,52
Hotel	0,27	**0,29**	0,27	0,50	**0,52**	0,45
Restaurante	0,19	**0,22**	0,20	0,52	**0,54**	0,50
Taxi	0,32	0,35	**0,39**	0,47	**0,51**	0,50
Comboio	0,30	**0,51**	0,32	0,65	**0,72**	0,56
Média macro	0,27	0,35	0,31	0,52	0,56	0,51
Média ponderada	0,26	0,32	0,29	0,54	0,58	0,50

Tabela 11: Desempenho do DST no MultiWOZ-PT, com reconhecimento de intenções integrado e diferentes métodos de pós-processamento.

em bruto (por exemplo, com intenções ou *slots*), facilitando assim a identificação de tendências e problemas de comunicação (por exemplo, numa pós-análise).

Agradecimentos

Este trabalho foi parcialmente apoiado pelo Plano de Recuperação e Resiliência (PRR) Português através do projeto C645008882-00000055, Centro para a IA Responsável; pela FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito do projeto CISUC – UID/CEC/00326/2020 e pelo Fundo Social Europeu, através do Programa Operacional Regional Centro 2020.

Referências

Budzianowski, Paweł, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan & Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5016–5026. 10.18653/v1/D18-1547

Carvalho, Nuno Ramos, Alberto Simões & José João Almeida. 2021. Bootstrapping a dataset and model for question-answering in Portuguese. Em *10th Symposium on Languages, Applications and Technologies (SLATE)*, 18:1–18:5. 10.4230/OASIcs.SLATE.2021.18

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 4171–4186. 10.18653/v1/N19-1423

Ding, Bosheng, Junjie Hu, Lidong Bing, Sharifah Mahani Aljunied, Shafiq Joty, Luo Si & Chunyan Miao. 2021. GlobalWoZ: Globalizing multiwoz to develop multilingual task-oriented dialogue systems. ArXiv [cs.CL]. 10.48550/arXiv.2110.07679

Fonseca, E, L Santos, Marcelo Criscuolo & S Aluísio. 2016. ASSIN: Avaliação de similaridade semântica e inferencia textual. Em *Computational Processing of the Portuguese Language (PROPOR)*, 13–15.

Freitas, Cláudia, Paula Carvalho, Hugo Gonçalo Oliveira, Cristina Mota & Diana Santos. 2010. Second HAREM: advancing the state of the art of named entity recognition in Portuguese. Em *International Conference on Language Resources and Evaluation (LREC)*,

Gao, Shuyang, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung & Dilek Hakkani-Tur. 2019. Dialog state tracking: A neural reading comprehension approach. ArXiv [cs.CL]. 10.48550/arXiv.1908.01946

Gonçalo Oliveira, Hugo, André Clemêncio & Ana Alves. 2020. Corpora and baselines for humour recognition in Portuguese. Em *12th Language Resources and Evaluation Conference (LREC)*, 1278–1285.

Gu, Xiaodong, Kang Min Yoo & Jung-Woo Ha. 2021. DialogBERT: Discourse-aware response generation via learning to recover and rank utterances. Em *AAAI Conference on Artificial Intelligence*, 12911–12919. 10.1609/aaai.v35i14.17527

He, Pengcheng, Xiaodong Liu, Jianfeng Gao & Weizhu Chen. 2020. DeBERTa: Decoding-enhanced bert with disentangled attention. ArXiv [cs.CL]. 10.48550/arXiv.2006.03654

Heck, Michael, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin,

Marco Moresi & Milica Gasic. 2020. TripPy: A triple copy strategy for value independent neural dialog state tracking. Em *21st Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 35–44. 10.18653/v1/2020.sigdial-1.4

Lamanov, Dmitry, Pavel Burnyshev, Katya Artemova, Valentin Malykh, Andrey Bout & Irina Piontkovskaya. 2022. Template-based approach to zero-shot intent recognition. Em *15th International Conference on Natural Language Generation*, 15–28. 10.18653/v1/2022.inlg-main.2

Lee, Chia-Hsuan, Hao Cheng & Mari Ostendorf. 2021. Dialogue state tracking with a language model using schema-driven prompting. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4937–4949. 10.18653/v1/2021.emnlp-main.404

Mota, Cristina & Diana Santos (eds.). 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo HAREM*. Linguateca.

Nekvinda, Tomáš & Ondřej Dušek. 2021. Shades of BLEU, flavours of success: The case of MultiWOZ. Em *1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, 34–46. 10.18653/v1/2021.gem-1.4

Ouyang, Yawen, Zhen Wu, Xinyu Dai, Shujian Huang & Jiajun Chen. 2022. Towards multi-label unknown intent detection. Em *29th International Conference on Computational Linguistics (COLING)*, 626–635.

Pais, Francisco, Patrícia Ferreira, Catarina Silva, Ana Alves & Hugo Gonçalo Oliveira. 2024. Question answering for dialogue state tracking in Portuguese. Em *16th International Conference on Computational Processing of Portuguese (PROPOR)*, 461–471.

Querido, Andreia, Rita Carvalho, João Rodrigues, Marcos Garcia, João Silva, Catarina Correia, Nuno Rendeiro, Rita Valadas Pereira, Marisa Campos & António Branco. 2017. LX-LR4DistSemEval: A collection of language resources for the evaluation of distributional semantic models of Portuguese. *Revista da Associação Portuguesa de Linguística* 3. 265–283. 10.26334/2183-9077/rapln3ano2017a15

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li & 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.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev & Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. ArXiv [cs.CL]. 10.48550/arXiv.1606.05250

Rastogi, Abhinav, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta & Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The Schema-Guided Dialogue Dataset. Em *AAAI Conference on Artificial Intelligence*, vol. 34 05, 8689–8696. 10.1609/aaai.v34i05.6394

Ribeiro, Eugénio, Ricardo Ribeiro & David Martins de Matos. 2019. Reconhecimento de actos de diálogo hierárquicos e multi-etiqueta em dados em Espanhol. *Linguamática* 11(1). 17–40. 10.21814/lm.11.1.278

Rodrigues, João, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso & Tomás Osório. 2023. Advancing neural encoding of Portuguese with transformer Albertina PT-*. ArXiv [cs.CL]. 10.48550/arXiv.2305.06721

Siddique, AB, Fuad Jamour & Vagelis Hristidis. 2021. Linguistically-enriched and context-awarezero-shot slot filling. Em *The Web Conference (WWW)*, 3279–3290. 10.1145/3442381.3449870

da Silva, Fábio Ricardo Araújo. 2018. *Detecção de ironia e sarcasmo em língua Portuguesa: Uma abordagem usando deep learning*: Universidade Federal de Mato Grosso. Tese de Mestrado. 10.13140/RG.2.2.18896.81924

Silva, Maurício. 2012. *O novo acordo ortográfico da língua portuguesa: o que muda, o que não muda*. Editora Contexto

Souza, Fábio, Rodrigo Nogueira & Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. Em *Intelligent Systems*, 403–417. 10.1007/978-3-030-61377-8_28

Tavares, Diogo, Pedro Azevedo, David Semedo, Ricardo Sousa & João Magalhães. 2023. Task conditioned BERT for joint intent detection and slot-filling. Em *EPIA Conference on Artificial Intelligence*, 467–480. 10.1007/978-3-031-49008-8_37

Williams, Jason D, Antoine Raux & Matthew Henderson. 2016. The dialog state tracking

challenge series: A review. *Dialogue & Discourse* 7(3). 4–33. /10.5087/dad.2016.301

Wu, Shijie & Mark Dredze. 2020. Are all languages created equal in multilingual BERT? ArXiv [cs.CL]. 10.48550/arXiv.2005.09093

Zang, Xiaoxue, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang & Jindong Chen. 2020. MultiWOZ 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. Em *2nd Workshop on Natural Language Processing for Conversational AI*, 109–117. 10.18653/v1/2020.nlp4convai-1.13

Zhang, Jian-Guo, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wan, Philip S Yu, Richard Socher & Caiming Xiong. 2019. Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. ArXiv [cs.CL]. 10.48550/arXiv.1910.03544

Zhang, Zheng, Ryuichi Takanobu, Qi Zhu, MinLie Huang & XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences* 63(10). 2011–2027. 10.1007/s11431-020-1692-3