Proposta recebida em Abril 2021 e aceite para publicação em Novembro 2021.

AIA-BDE: um Corpo de Perguntas, Variações e outras Anotações

AIA-BDE: a corpus of Portuguese Questions, Variations and other Annotations

Hugo Gonçalo Oliveira
CISUC, DEI, Universidade de Coimbra

Ana Alves
CISUC, Universidade de Coimbra
ISEC, Instituto Politécnico de Coimbra

Resumo

Apresentamos neste artigo o corpo AIA-BDE, que tem como principal objetivo a avaliação de sistemas que procuram associar necessidades de informação expressas em linguagem natural a perguntas com resposta conhecida (i.e., FAQ). Este corpo inclui várias perguntas no domínio da Administração Pública em Portugal e respetivas respostas. A 855 dessas perguntas foram adicionadas, manual e automaticamente, formas alternativas de as fazer, a que chamamos variações, e que podem ser utilizadas para simular interações de humanos. Essas perguntas encontram-se classificadas de acordo com a sua origem, com quatro valores possíveis, e têm ainda associado um tipo, atribuído com base na opinião de cinco anotadores. Para além de apresentar o AIA-BDE, ilustramos como pode ser utilizado através de três experiências, com resultados que podem ser vistos como base para melhorias futuras: associação de variações às respetivas perguntas; identificação automática da origem das variações; e classificação automática das perguntas quanto ao seu tipo.

Palavras chave

corpora, FAQs, resposta a perguntas, paráfrases, similaridade semântica, classificação de texto

Abstract

We present the AIA-BDE corpus, which has as main goal the evaluation of computational systems that attempt at assigning questions with known answers (i.e., FAQs) to information needs, expressed in natural language. This corpus includes several questions in the domain of the Portuguese Public Administration and their answers. To 855 of those questions, alternative ways of making them were manually and automatically added. We call them variations and they can be used in the simulation of human user interactions. Such questions are classified according to their source, with four possible values, and have also a question type, based on the opinion of five human annotators. Besides presenting AIA-BDE, we illustrate how it can be used through three experiments, with results that might be seen as the baselines for future improvements, namely: variation assignment to the original questions; automatic automatic identification of the questions according to their source; and automatic classification of the questions according to their type.

Keywords

copora, FAQs, question answering, paraphrases, semantic similarity, text classification

1. Introdução

O projeto *AIA: Apoio Inteligente a Empreendedores* decorreu entre os anos de 2018 e 2020 numa colaboração entre o Centro de Informática e Sistemas da Universidade de Coimbra (CISUC), o INESC-ID, e a Agência para a Modernização Administrativa (AMA), e teve como objetivo inicial o estudo de mecanismos de interação em linguagem natural para assistir automaticamente empreendedores em Portugal. Nesse contexto, foram exploradas diferentes técnicas para agentes artificiais com a capacidade de resposta automática a perguntas colocadas em português (Gonçalo Oliveira et al., 2019; Santos et al., 2020b,a). A maioria destes agentes usa como base listas de Perguntas Já Respondidas (em inglês, *Frequently Asked Questions*) e procura associar as perguntas do utilizador a perguntas conhecidas.

Para avaliar progressos, ao longo do projeto foi feita uma recolha de dados, focada em FAQ no domínio do problema, inicialmente a partir de conteúdos do Balcão do Empreendedor (BDE), desde 2019 integrado no portal e-Portugal[1]. O resultado desta recolha de perguntas, da sua expansão e da sua organização, é o corpo AIA-BDE, cuja versão 2.1 apresentamos aqui.

Apesar desta versão incluir um conjunto adicional de perguntas e respostas, a principal con-

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

DOI: 10.21814/lm.13.2.350

tribuição do trabalho está num conjunto de 855 perguntas e respetivas respostas, obtidas a partir de quatro fontes distintas, e num total de 5.298 variações, algumas produzidas automaticamente, outras manualmente, por diferentes grupos de pessoas. As variações são outras formas de fazer as mesmas perguntas, utilizando outras palavras ou construções sintáticas alternativas, na maior parte dos casos paráfrases das perguntas originais. Considerando que um agente que dá respostas com base em perguntas conhecidas (i.e., FAQ) deve ter a capacidade de lidar com perguntas feitas de diferentes formas, a principal utilidade destas variações está na simulação de interações humanas e, consequentemente, na avaliação de agentes deste tipo.

Para além das variações, numa fase final do projeto, foi atribuída uma (ou mais) de nove classes a cada uma das 855 perguntas, de acordo com o seu tipo. Ainda que, até à data, esta anotação tenha sido pouco explorada, a identificação automática do tipo de pergunta pode ser vista como um problema de classificação, com utilidade para sistemas de resposta automática a perguntas, dado que esse tipo vai condicionar a resposta dada.

Este artigo descreve o corpo AIA-BDE, que acreditamos tratar-se de um contributo importante para o desenvolvimento e avaliação de sistemas de resposta automática a perguntas, deteção de paráfrases ou similaridade semântica textual, entre outros, na língua portuguesa. É também por isso que disponibilizamos o corpo a todos os potenciais interessados.

Antes de descrever o corpo, apresentamos algum trabalho relacionado, mais propriamente outros corpos com caraterísticas semelhantes, para português e para outras línguas, bem como a sua utilização. Depois de descrever o conteúdo do AIA-BDE, relatamos três experiências realizadas com o corpo, nomeadamente na associação automática de variações às perguntas originais; classificação automática de perguntas de acordo com a sua origem; e classificação automática do tipo de pergunta. Os resultados obtidos podem ser vistos como base para experiências futuras com o corpo, onde outras técnicas podem vir a ser exploradas.

2. Trabalho Relacionado

Resposta Automática a Perguntas (RAP) é uma tarefa que tem como objetivo obter, de forma automática, respostas para perguntas colocadas em linguagem natural. Ainda que esse não seja um requisito, sistemas de RAP são normalmente focados em perguntas factuais (Voorhees, 2008) e baseiam-se em técnicas de Recuperação de Informação (IR, do inglês *Information Retrieval*) (Kolomiyets & Moens, 2011) para encontrar a resposta em coleções de documentos.

O desenvolvimento de sistemas de RAP em português foi impulsionado pelo fórum de avaliação CLEF, onde esta língua esteve presente de 2004 (Magnini et al., 2004) a 2008 (Forner et al., 2008). Neste âmbito, destaca-se a criação da coleção CHAVE (Santos & Rocha, 2004), que inclui textos noticiosos publicados entre 1994 e 1995 e que foi usada para avaliações de IR e RAP (QA@CLEF).

Para avaliação de IR, foram definidos tópicos apropriados, considerando que podiam ser usados em pesquisas noutras línguas, e documentos foram marcados como relevantes ou não para cada tópico. A avaliação de RAP é semelhante, mas com perguntas factuais em vez de tópicos, incluindo a indicação do tipo de resposta esperada, e com respostas em vez de documentos. Ao longo das várias edições onde o português esteve incluído, foi criado um total de 4.380 perguntas em português e respetivas respostas na coleção CHAVE, quando essa resposta existe.

Outra avaliação relacionada foi o Págico (Mota et al., 2012), para a qual foi criada uma lista de 150 tópicos acerca da cultura lusófona e, para cada um, indicadas as páginas da Wikipédia que lhes respondiam, sendo que os tópicos correspondiam a perguntas cuja resposta, regra geral, não se encontrava em apenas uma página da Wikipédia.

A resposta automática a Perguntas Já Respondidas (doravante, FAQ) é uma tarefa específica no âmbito de IR e RAP, em que para cada pergunta há uma resposta bem definida. Devido à sua natureza e estrutura, listas de FAQ têm sido exploradas no desenvolvimento de sistemas de esclarecimento de dúvidas sobre um determinado domínio, em alguns casos culminando com a disponibilização das coleções usadas.

Um dos primeiros sistemas deste tipo, para a língua inglesa, terá sido o FAQFinder (Burke et al., 1997), que para além de uma abordagem baseada em IR, tira partido de conhecimento semântico na base de conhecimento lexical WordNet (Fellbaum, 1998). Para a língua croata, um sistema do mesmo tipo foi treinado e testado para, com base em técnicas de similaridade semântica textual (Agirre et al., 2012), fornecer respostas com base em 1.222 perguntas recolhidas a partir do sítio de uma operadora móvel na Web (Karan et al., 2013). As perguntas e respostas foram ainda enriquecidas por

10 voluntários que, sem conhecer à partida as perguntas recolhidas, criaram questões-exemplo que poderiam ser colocadas por utilizadores reais, assim como as suas paráfrases. O último passo consistiu na aplicação de modelos tradicionais de IR (e.g., pesquisa com base em palavras-chave, TF-IDF, modelação de linguagem) para, por cada questão colocada pelos voluntários, recuperar FAQ cuja relevância binária foi finalmente atribuída de forma manual.

A coleção FAQIR (Karan & Šnajder, 2016) tem 4.313 FAQ acerca de "manutenção e reparações", obtidas a partir do portal *Yahoo! Answers*, e 1.233 perguntas de utilizadores acerca desse domínio, obtidas através do parafraseamento de 50 necessidades de informação base. A cada uma das perguntas anteriores foi posteriormente associado um de quatro valores de relevância para cada FAQ, com base em métodos de IR não supervisionados: Relevante (associação perfeita), Útil (tópico relacionado e útil), Inútil (tópico relacionado, mas sem utilidade), e Irrelevante (tópico não relacionado). Contudo, apesar do elevado número de FAQ, apenas uma pequena parte (≈22%, 779) tem pelo menos uma pergunta considerada relevante ou útil.

Os mesmos autores criaram uma nova coleção de FAQ, agora no domínio das aplicações web (Karan & Šnajder, 2018), obtida a partir do portal StackExchange[2]. A recolha focou-se nas perguntas mais populares neste domínio, correspondentes a 125 necessidades de informação. Ao associar cada uma dessas perguntas a diferentes respostas dadas por utilizadores e consideradas boas, de acordo com os votos dados, foi possível obter 719 FAQ. Ainda que as perguntas fossem constituídas por um título e uma descrição mais detalhada, foram usados apenas os títulos, após uma revisão manual que procurou garantir que continham toda a informação relevante. Depois desta recolha, dois anotadores produziram, cada um, cinco formas diferentes de exprimir a necessidade de informação associada a cada FAQ. Este processo resultou em 1.250 variações e à associação da relevância binária de cada forma para cada FAQ.

Ainda a este respeito, a QA4FAQ (Caputo et al., 2016) foi um tarefa de avaliação conjunta para resposta automática a perguntas baseada em FAQ escritas em italiano, organizada no contexto da avaliação EVALITA 2016. O objetivo consistia em recuperar FAQ relevantes para perguntas feitas por utilizadores. Neste contexto, foram disponibilizadas 406 FAQ (perguntas, respostas, etiquetas), 1.232 perguntas de utilizadores recolhidas a partir de registos de um sistema IR, e um conjunto de mapeamentos entre perguntas e FAQ.

Uma tarefa relacionada com a resposta a FAQ é a resposta a perguntas da comunidade (em inglês, *Community Question Answering*) (Nakov et al., 2015, 2016, 2017), onde o objetivo é ordenar pares de perguntas-perguntas e perguntas-comentários em fóruns de discussão na Web, de acordo com a sua similaridade. O fórum do portal *Qatar Living*[3] tem sido utilizado como fonte de dados para esta tarefa. A partir de uma lista de perguntas originais, outras perguntas relacionadas são obtidas para cada, assim como os primeiros comentários nos seus tópicos. A relevância de cada pergunta relacionada para a pergunta original foi atribuída, e o mesmo foi feito para a relevância dos comentários em relação às perguntas originais e à pergunta original, embora diferentes anotações tenham sido utilizadas em diferentes sub-tarefas. As perguntas foram finalmente geradas a partir do assunto de cada pergunta original e o motor de busca Google foi usado para recolher 200 comentários de perguntas no sítio do fórum. Os resultados com dez ou mais comentários e perguntas com menos de 2.000 caracteres foram consideradas como perguntas relacionadas válidas.

No que concerne especificamente ao português, e tanto quanto conhecemos, o mais próximo com uma coleção de FAQ e respetivas respostas será a coleção MilkQA (Criscuolo et al., 2017), que inclui perguntas colocadas de forma mais densa, no domínio dos laticínios, seguidas pelas suas respostas. A disponibilização do corpo AIA-BDE é mais uma contribuição neste sentido, já que inclui FAQ em português, no domínio da administração pública, e está preparada para avaliar um conjunto de tarefas relevantes para sistemas de IR, RAP, e mesmo diálogo, com foco na associação de interações em linguagem natural com perguntas conhecidas.

Por se abordar normalmente como uma tarefa de IR, o desempenho de sistema de RAP é avaliado com recurso a medidas como a precisão, a abrangência e a medida-F. Já na recuperação de FAQ é normal assumir que existe uma e uma só resposta para cada pergunta (Burke et al., 1997), e, nesse caso, o desempenho é dado pela proporção de perguntas de uma lista para as quais foi dada a resposta correta (Burke et al., 1997; Caputo et al., 2016). Como a mesma pergunta pode ser feita de diferentes formas, o principal desafio é associar qualquer pergunta do utilizador a uma das perguntas conhecidas. Aqui é co-

[2] http://www.stackexchange.com/

[3] https://www.qatarliving.com/

mum recorrer a técnicas de IR, podendo passar-se pelo reconhecimento automático de paráfrases ou cálculo de similaridade semântica textual. Para as últimas tarefas, foram também organizadas avaliações conjuntas em português (Fonseca et al., 2016; Real et al., 2020), que resultaram na disponibilização de coleções que incluem pares de frases e respetivos valores de similaridade e classificação de paráfrases.

Um outro tipo de RAP em que o interesse tem aumentado recentemente é a resposta extrativa a perguntas (em inglês, *Extractive Question Answering*). Este interesse tem sido impulsionado pela sua inclusão na avaliação de modelos de compreensão de linguagem natural (em inglês, *Machine Reading Comprehension*) (Devlin et al., 2019), recorrendo a coleções como a SQuAD (Rajpurkar et al., 2016), que inclui mais de 100 mil perguntas colocadas acerca de artigos da Wikipédia, e respetivas respostas. Apesar de originalmente em inglês, foram recentemente disponibilizadas duas traduções do SQuAD para português (Carvalho et al., 2021; Guillou, 2021).

Tal como acontece nas coleções para RAP e de FAQ, no SQuAD, a ordem das perguntas para o mesmo documento não é importante. Por outro lado, há corpos com uma estrutura semelhante ao SQuAD que procuram simular conversas, e por isso mais adequados, por exemplo, ao desenvolvimento e avaliação de sistemas de diálogo. Aqui destacamos o QuAC (Choi et al., 2018) e o CoQA (Reddy et al., 2019), ambos criados com base numa tarefa em que uma pessoa realiza perguntas acerca de um dado assunto e outra responde, tão naturalmente quanto possível, tendo por base um texto acerca do mesmo assunto. Já no âmbito dos sistemas de diálogo orientados à resolução de tarefas, há corpos como o MultiWOZ (Budzianowski et al., 2018) que podem ser utilizados para treino e avaliação. Este corpo tem vários diálogos com mais do que um turno, estruturados em atos de diálogo (em inglês, *Dialog Acts*) atribuídos manualmente, que incluem uma intenção (em inglês *intent*) e um conjunto de pares atributo-valor de acordo com a tarefa a realizar. Um exemplo será o ato de *informar(domínio=hotel, preço=moderado)*, que representa a intenção de obter nomes de hotéis com um preço moderado.

Ainda no âmbito do diálogo, vários corpos têm sido utilizados no desenvolvimento de sistemas que seguem uma abordagem orientada a dados (em inglês, *data-driven*), mais propriamente, para aprenderem a traduzir interações do utilizador em respostas do sistema. Aqui destacam-se diálogos obtidos a partir de plataformas de conversação ou redes sociais, respetivamente compilados no Ubuntu Dialogue Corpus (Lowe et al., 2015) ou no Twitter Conversation Corpus (Ritter et al., 2011), ou diálogos obtidos a partir de legendas de filmes, compilados, por exemplo, nos corpos Open Subtitles (Lison & Tiedemann, 2016).

Ao contrário dos sistemas de recuperação de FAQ, a avaliação de sistemas de diálogo aberto (e.g., Vinyals & Le (2015)) é um desafio, e passa muitas vezes por comparar as respostas de um dado sistema com as respostas dadas por humanos, no mesmo contexto (Gunasekara et al., 2020). Na medida em que não existem objetivos claramente definidos para esses sistemas, o problema mantém-se mesmo quando são usadas métricas recentes, como a BertScore (Zhang et al., 2020), que considera representações semânticas e não se limita a comparar a sobreposição de sequências.

Uma alternativa mais demorada e dispendiosa passa por colocar utilizadores humanos a interagir com um dado sistema e avaliarem o quão natural e fluído decorreu o diálogo (Gunasekara et al., 2020). Quando estão a ser avaliados sistemas orientados à realização de tarefas, o sucesso na concretização da tarefa poderia ser acrescentado a esta avaliação (Wen et al., 2017). O recrutamento de utilizadores, em qualquer um dos casos, pode ser feito via *crowdsourcing*. Porém, para medir o progresso do sistema, a avaliação deveria ser realizada para cada nova atualização.

Especificamente para o português, os corpos de legendas de filmes têm sido usados por agentes conversacionais, não para responder a perguntas ou resolver tarefas em concreto, mas para melhor lidar com interações fora do domínio (Magarreiro et al., 2014). Considerando que, para decidir que tipo de resposta dar e como a obter é importante identificar o tipo de interação, relacionado com o ato de diálogo, para o português há também a coleção UC-PT (Fernandes et al., 2019), que inclui interações em diálogos com um conjunto de anotações: pergunta ou não; pergunta com resposta sim/não; pergunta de cariz pessoal.

3. Conteúdo do Corpo

Esta secção descreve o formato e o conteúdo do corpo AIA-BDE. Começa por explicar a estrutura dos ficheiros em que é distribuído, e depois apresenta e quantifica a origem das perguntas, as variações e respetiva produção e, finalmente, a atribuição do tipo de pergunta. Nas secções seguintes, ilustram-se algumas das possibilidades que o AIA-BDE oferece através de um conjunto de experiências realizadas.

```
S:Espaço Empresa
...
P:Como pedir o Cartão Provisório de Identificação de Pessoa Coletiva?
VG1:Como solicitar o cartão de identidade provisório?
VG2:Como solicitar o cartão de identidade provisório?
VUC:Como posso obter o cartão provisório de identificação de pessoa coletiva?
VUC:Onde posso pedir o cartão provisório de pessoa coletiva?
VIN:Como posso pedir o Cartão provisório de identificação de pessoa coletiva?
VIN:Qual o procedimento para obter o Cartão Provisório de Identificação de Pessoa
Coletiva?
R:O Cartão Provisório de Identificação de Pessoa Coletiva deixou de ser emitido,
(...)  Atualmente, existe apenas o Cartão da Empresa e o Cartão de Pessoa Coletiva,
que são emitidos para entidades definitivamente registadas ou inscritas.
```

Figura 1: Sequência de linhas do AIA-BDE 2.1

3.1. Estrutura e Organização

O corpo AIA-BDE é distribuído a partir do repositório https://github.com/NLP-CISUC/AIA-BDE, num ficheiro principal, `AIA-BDE_v2.1.txt`, e de outros ficheiros complementares. Todos os ficheiros são constituídos por perguntas e respetivas respostas. Mais propriamente, cada linha começa com um marcador que pode indicar se se trata de uma pergunta (`P:`), da resposta à pergunta anterior (`R:`), ou ainda a origem de todas as perguntas que se seguem (`S:`).

Especificamente no ficheiro principal, entre as 855 perguntas e as respetivas respostas, encontram-se variações da pergunta, cujos marcadores começam por `V`. A título de exemplo, a Figura 1 mostra uma sequência de linhas deste ficheiro. Na versão 2.1, este ficheiro inclui perguntas obtidas a partir de quatro fontes, mais propriamente:

- 625 perguntas do *Espaço Empresa* (EE), que cobrem informações relacionadas com o exercício de uma atividade económica e com o ciclo de vida de uma empresa;
- 118 do *Guia de Aplicação do Regime Jurídico de Acesso e Exercício de Atividades de Comércio, Serviços e Restauração* (RJACSR);
- 56 acerca da legislação de *Alojamento Local* (AL);
- 56 obtidas a partir de um conjunto de guias relacionados com *Apoios sociais* (AS), as últimas acrescentadas.

Ao nível das origens, o corpo tem uma distribuição altamente desequilibrada, com mais de três quartos das perguntas originárias do EE. De forma a não perder a informação relativa à origem mais específica destes documentos, foram introduzidos dois marcadores com a indicação da zona (`SS`) e do ficheiro (`SSS`) onde estas perguntas se encontravam dentro do EE.

O ficheiro `AIA-BDE_tipo_pergunta.txt` tem exatamente as mesmas perguntas e respostas que o anterior, mas: (i) não inclui variações; e, (ii) depois de cada resposta, tem uma linha iniciada por um marcador `F:` seguida do nome do tipo ou tipos da pergunta anterior, separados por vírgulas, com a respetiva confiança associada a cada, separada por um `#` (e.g., `Condição#0.6,Procedimento#0.4`).

Para além dos dois ficheiros anteriores, há um conjunto de ficheiros com mais perguntas, respostas e, em alguns casos, variações, todas do mesmo tipo e produzidas automaticamente a partir de documentos estruturados. Estas perguntas foram geradas no âmbito do projeto AIA e, mesmo nos casos em que existem variações, elas seguem todas uma estrutura padrão. A sua geração teve como principal objetivo uma rápida ampliação do número de perguntas a que os agentes desenvolvidos poderiam responder. No entanto, devido à sua simplicidade e falta de revisão manual, a sua utilidade para fins de avaliação é limitada, e não foram por isso utilizados em nenhuma das experiências descritas neste artigo.

Entre os ficheiros anteriores destacamos o `AIA_actividades.txt` e o `AIA_licencas.txt`. O primeiro inclui 844 perguntas que usam uma de três formas para perguntar o que é determinada atividade económica, as outras duas formas como variação (marcada por `VAU`), e como resposta uma definição da atividade. O segundo ficheiro inclui 1.281 perguntas de um de dois tipos: (i) o que permite determinada licença; ou (ii) licença necessária para fazer qualquer coisa. Cada uma também pode ser feita de uma de duas formas, sendo a outra usada como variação. A Figura 2 ilustra estes dois ficheiros com uma seleção de linhas de cada um.

```
P:O que é um Café?
VAU:O que faz um Café?
VAU:Para que serve um Café?
R:Estabelecimentos de bebidas que servem, através de pagamento, bebidas e cafetaria ...
```

```
P:De que preciso para circular, parar e estacionar veículos de tração animal?
VAU:O que permite circular, parar e estacionar veículos de tração animal?
R:Veículo de tração animal - licença de circulação
P:O que permite a licença Grua - licença de ocupação do espaço público ?
VAU:O que posso fazer com uma Grua - licença de ocupação do espaço público ?
R:Permite a instalação de uma grua (aparelho para levantar e deslocar corpos pesados), ...
```

Figura 2: Linhas dos ficheiros `AIA_actividades.txt` (cima) e `AIA_licencas.txt` (baixo).

3.2. Perguntas e Variações

Para cada uma das 855 perguntas no ficheiro principal do AIA-BDE foram produzidas variações, isto é, reformulações da pergunta original utilizando outras palavras ou construções, mas mantendo o significado original ou um suficientemente próximo, de tal forma que a resposta original continue a ser válida.

A sua produção teve em conta que, na maioria dos casos, os utilizadores não escrevem uma pergunta exatamente da mesma forma que ela se encontra numa lista de perguntas já respondidas (FAQ). Isto implica que, para ter sucesso, um sistema computacional que procure encontrar respostas com base numa lista de perguntas, terá de conseguir associar perguntas com base na sua proximidade semântica, ainda que feitas por outras palavras. Ou seja, terá de lidar com fenómenos como a sinonímia e o relacionamento semântico ou, ao nível da frase, similaridade semântica textual (Agirre et al., 2012), parafraseamento e inferência (Bowman et al., 2015). Assim, as variações têm como objetivo principal permitir a avaliação de sistemas de resposta automática a perguntas com resposta conhecida, ou simplesmente de sistemas focados nas tarefas anteriores, mas em contexto interrogativo.

Ainda que produzidas de diferentes formas, para cada pergunta do AIA-BDE há pelo menos cinco variações. Por não haver uma forma ideal de produzir variações, e porque a sua criação manual é um processo moroso, as variações foram sendo produzidas ao longo do tempo, por diferentes pessoas, e seguindo abordagens diferentes. Assim, optamos por marcá-las consoante o processo de criação, em alguns casos automático e noutros manual. Para as primeiras, utilizamos a API do Google Translate API[4] como uma abordagem rápida e de baixo custo para gerar paráfrases da pergunta original. Mais propriamente, cada pergunta tem duas variações deste tipo: tradução do texto em português para inglês e novamente para português (VG1), e tradução do resultado anterior novamente para inglês e para português (VG2). Dada a simplicidade da abordagem, algumas das variações acabam por ser muito próximas, ou até iguais, à pergunta original. Mais propriamente, há 51 variações VG1 e 41 VG2 iguais à pergunta original. Numa minoria de casos, a semântica acaba mesmo por sofrer alterações, devido a problemas na tradução e consequente introdução de termos incorretos.

Devido às limitações da abordagem anterior, foram também produzidas variações de forma manual. Neste caso, por terem sido criadas por diferentes grupos de pessoas, optamos por separá-las em três tipos:

- Variações produzidas pela equipa de investigadores do projeto AIA na Universidade de Coimbra (VUC);
- Variações produzidas por alunos da unidade curricular de Língua Natural, leccionada em mestrados do Instituto Superior Técnico (VIN);
- Variações produzidas com recurso à plataforma de *crowdsourcing* Amazon Mechanical Turk[5] (VMT).

Para qualquer um dos tipos, foi pedido aos voluntários que lessem tanto a pergunta original como a sua resposta, e para reescrever a pergunta usando outras palavras, ainda que mantendo o significado original ou, pelo menos, um significado próximo ou implicado pelo original.

A Tabela 1 contabiliza, para cada origem, o número de perguntas e variações de cada tipo disponível. Apresenta ainda o número médio de átomos nas perguntas de cada origem (**Comprimento**), que permite verificar que as perguntas de AL são, regra geral, mais longas,

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

[5] https://www.mturk.com/

Origem	Perguntas	Comprimento	Variações					Total
			VG1	VG2	VUC	VIN	VMT	
EE	625	11,6±5,8	625	625	430	2.279	0	4.584
RJACSR	118	14,6±9,1	118	118	380	0	0	734
AL	56	20,1±12,5	56	56	122	0	0	290
AS	56	12,5±5,0	56	56	0	0	168	336
Total	855	12,6±7,3	855	855	932	2.279	168	5.944

Tabela 1: Distribuição de perguntas e variações por origem.

e as do EE mais curtas. Para cada pergunta original há uma variação do tipo VG1 e VG2, no entanto, a existência dos restantes tipos é variável, perfazendo pouco mais de 5.000 variações. O tipo mais predominante são as variações VIN, no entanto, estas foram realizadas apenas para perguntas do EE. Por outro lado, as variações VMT foram produzidas apenas para as perguntas AS, mais propriamente, três para cada pergunta, e não há mais nenhum tipo de variação manual para estas perguntas.

A Tabela 2 ilustra o conteúdo do corpo AIA-BDE com uma pergunta para cada origem, seguida de um conjunto de variações e, finalmente, da resposta. No primeiro exemplo, do EE, as variações VIN aparentam ser mais "conservadoras" do que as VUC, o que se pode ver neste e nos restantes exemplos. Isto não acontece apenas aqui e é uma das razões para termos decidido marcar as variações de acordo com a forma de criação. Podemos considerar que na criação das VUC terá havido mais criatividade, com maior utilização de sinónimos e variações a omitir partes da pergunta original. Finalmente, e apesar das guias serem as mesmas, nas VMT houve um menor controlo no processo de criação, o que torna as suas diferenças para a pergunta original mais variáveis. Veja-se o último exemplo da tabela.

3.3. Tipo de pergunta

A última anotação adicionada às perguntas do corpo AIA-BDE foi o tipo de pergunta. Apesar de não ter sido explorada no contexto do projeto AIA, a sua identificação automática pode ser útil para um sistema computacional saber como responder ou onde procurar pela resposta. Num sistema de diálogo, o tipo de pergunta estará relacionado com os atos do diálogo e, por essa razão, ser útil para o seu reconhecimento.

Depois de olhar para as várias perguntas do AIA-BDE, foram definidas nove tipos em que as perguntas poderiam ser classificadas, todas elas com exemplos identificados. Esses tipos foram: Binário, Condição, Custo, Definição, Local, Pré-requisito, Procedimento, Tempo, Vantagem.

De forma a agilizar o processo de anotação dos tipos e a considerar mais do que uma opinião nesta escolha, optámos por recorrer à plataforma de *crowdsourcing* Amazon Mechanical Turk. Mais propriamente, o tipo de cada pergunta foi atribuído de forma independente por cinco trabalhadores.

Antes de realizar a tarefa, foram apresentadas diretivas onde se descrevia cada um dos tipos e se incluía um exemplo para cada um. A combinação das cinco anotações permitiu calcular a confiança relativamente à adequação de cada tipo a cada pergunta. Mais propriamente, para uma pergunta, a confiança num tipo resulta da divisão do número de vezes que esse tipo foi atribuído pelo total de anotações obtidas (5).

A Tabela 3 apresenta a distribuição de perguntas por tipo e confiança. No ficheiro disponibilizado, optou-se por omitir os tipos com menor confiança nos casos em que o tipo com maior confiança tinha sido atribuído por mais de dois trabalhadores. Ou seja, quando há um tipo atribuído por três ou quatro trabalhadores e os outros por apenas um, apenas se apresenta o primeiro. Verifica-se que há tipos frequentemente atribuídos, nomeadamente a Definição e o Procedimento, enquanto outros são mais raros. Por exemplo, não há qualquer pergunta dos tipos Binário e Local com confiança 100%.

Na Tabela 4 apresentamos exemplos de perguntas do EE com os seus tipos e confianças calculadas. Tal como em alguns exemplos da tabela, há várias perguntas que acabam por ter mais do que um tipo e, na maior parte das vezes, estão ambos corretos e nem fará muito sentido escolher apenas um. Quando definimos os tipos preocupamo-nos mais em abranger todas as perguntas do que propriamente garantir que os tipos eram mutuamente exclusivos. A nossa opção por recorrer a cinco trabalhadores por pergunta e de incluir a confiança também está relacionada com esta situação. E assim, potenciais interessados poderão utilizar essa informação da forma que

Origem	Var	Texto
EE	P	*É necessário submeter documentos para efectivar o Registo por Depósito?*
	VG1	*É necessário enviar documentos para efetuar o registro por depósito?*
	VG2	*Preciso enviar documentos para registrar por depósito?*
	VIN	Para efectivar o Registo por Depósito é necessário submeter documentos?
	VIN	*Tenho que submeter documentos para efectivar o Registo por Depósito?*
	VUC	*De que documentos preciso para realizar um registo por depósito?*
	VUC	*Que documentos é necessário enviar para fazer um registo de depósito?*
	R	*Sim, deverá submeter os documentos que titulem o acto requerido.*
RJACSR	P	*Qual a coima aplicável às contraordenações graves?*
	VG1	*Qual é a multa aplicável à falta grave?*
	VG2	*Qual é a multa aplicável à falta grave?*
	VUC	*coima para contraordenação grave*
	VUC	*Qual o valor da multa para contraordenações graves?*
	R	*As contraordenações graves são sancionáveis com coima: ...*
AL	P	*No alojamento local é obrigatória a certificação energética? Em que termos deve ser efetuada?*
	VG1	*No alojamento local é obrigatório a certificação energética? Em que condições deveria ser feito?*
	VG2	*A certificação energética é necessária em alojamento local? Em que condições deve ser feito?*
	VUC	*Como deve ser feita certificação energética do meu alojamento local?*
	VUC	*Qual o procedimento para certificar energeticamente o meu alojamento local?*
	R	*De acordo com esclarecimento da DGEG (Direção-Geral de Energia e Geologia) ...*
AS	P	*Quando é que me dão uma resposta sobre o apoio social a crianças e jovens?*
	VG1	*Quando recebo uma resposta sobre apoio social para crianças e jovens?*
	VG2	*Quando recebo uma resposta sobre apoio social a crianças e jovens?*
	VMT	*Quando receberei a resposta do apoio social a crianças e jovens?*
	VMT	*Tenho que esperar muito para ter uma resposta sobre o apoio social a crianças e jovens?*
	R	*Depois de fazer a sua inscrição na instituição que lhe interessa, pode acontecer ter de ficar em lista de espera...*

Tabela 2: Exemplos de perguntas, variações e respostas no AIA-BDE.

Tipo	Confiança			
	100%	80%	60%	40%
Binário	0	8	35	66
Condição	1	12	45	124
Custo	15	11	5	8
Definição	115	64	38	58
Local	0	4	3	11
Pré-requisito	2	10	24	58
Procedimento	40	48	59	96
Tempo	23	14	9	15
Vantagem	2	9	8	5

Tabela 3: Distribuição de perguntas por tipo.

lhes for mais conveniente.

4. Associação de variações a perguntas

Esta e as próximas secções ilustram algumas das experiências que o AIA-BDE permite realizar. A primeira foca-se nas variações e, em trabalhos anteriores (Burke et al., 1997; Karan et al., 2013; Karan & Šnajder, 2018), foi apelidada de Recuperação de FAQ (*FAQ retrieval*). O objetivo passa por associar perguntas feitas de diferentes formas a perguntas conhecidas e já com uma resposta conhecida. Isto permite simular a capacidade de um sistema identificar paráfrases ou calcular a similaridade semântica entre frases interrogativas, um cenário com aplicabilidade, por exemplo, nos sistemas de diálogo ou de resposta a perguntas. Se a base de conhecimento de um sistema deste tipo for uma lista de FAQ, ao associar a pergunta recebida a uma conhecida, ele pode imediatamente retornar a sua resposta.

Pergunta	Tipo
O que é o Cartão da Empresa e o Cartão de Pessoa Coletiva?	Definição#0.8
Para a constituição de uma empresa através do serviço de criação de Empresa Online é necessária a presença de todos os sócios?	Procedimento#0.4, Pré-requisito#0.4
Como pedir o Cartão Provisório de Identificação de Pessoa Coletiva?	Procedimento#1
O Cartão de Identificação de Pessoa Coletiva ou entidade equiparada, emitido pelo RNPC e de que sou titular, continua a ser válido?	Procedimento#0.4, Binário#0.4
Quando é possível o levantamento do capital social da Empresa Online?	Procedimento#0.4, Tempo#0.4
Onde posso adquirir um certificado digital qualificado?	Procedimento#0.4, Local#0.6
O que acontece se o trabalhador adoecer durante as férias?	Condição#0.6
Quais as vantagens de aderir a um centro de arbitragem?	Vantagem#1
Onde posso pedir uma certidão permanente?	Local#0.8
Qual o custo do Cartão da Empresa e do Cartão de Pessoa Coletiva?	Custo#0.8
O Cartão da Empresa e o Cartão de Pessoa Coletiva podem ser cancelados?	Binário#0.8

Tabela 4: Seleção de perguntas do Espaço Empresa, respetivo tipo e confiança.

O que acabamos por avaliar com esta primeira experiência é um conjunto de métodos para o cálculo da similaridade semântica. Mais propriamente, para cada variação, utilizamos cada um dos métodos para calcular a similaridade com cada uma das 855 perguntas, e analisamos quantas vezes ele atribuiu a maior similaridade à pergunta original. Como esta análise pode ser limitada, olhamos ainda para o número de vezes em que a pergunta original está nas três (top-3) e nas cinco (top-5) mais similares. Isto porque, principalmente em casos de dúvida, pode ser mais útil fornecer três, ou até cinco respostas, em que uma delas é a pretendida, do que não fornecer nenhuma ou fornecer uma que não a desejada.

Foram testados diferentes métodos, todos eles não-supervisionados. Esta opção prende-se, por um lado, com a escassez de dados para treino, e, por outro, tem em vista a flexibilidade dos métodos e sua aplicabilidade a diferentes quantidades de dados, com diferentes origens (e.g., FAQ noutros domínios). A principal diferença entre os vários métodos testados é a forma adoptada para representar os textos. Aqui incluímos métodos mais simples, baseados em técnicas tradicionais de IR, e outros baseados na representação de texto com recurso a modelos distribucionais, nomeadamente *word embeddings* e *sentence embeddings*. Mais precisamente, testamos:

- Um método baseado numa biblioteca de IR para Python, Whoosh[6], com a configuração base (Whoosh base), ou com a análise de stems (StemmingAnalyzer) e tratamento de acentos (Charset Filter) ativados para português (Whoosh+). Para ambas as configurações, cada pergunta do AIA-BDE foi representada por um documento com dois campos —pergunta e resposta, com a pesquisa feita apenas no primeiro. O parâmetro *group* usou o valor *orGroup*, para evitar que todos os termos da pergunta fossem obrigatórios, e o método de ordenamento utilizado foi o BM25F.
- Métodos baseados na representação das perguntas através de *word embeddings* estáticos, nomeadamente word2vec CBOW (CBOW), GloVe e FastText. Mais propriamente, considerou-se que cada pergunta seria representada pelo vetor médio dos vetores dos seus átomos no modelo de *embeddings*. Para cada modelo, considerou-se também uma variação com uma média pesada, usando como peso o valor do TF-IDF[7] de cada átomo na pergunta em relação ao corpo constituído por todas as perguntas originais do AIA-BDE. Foram usados modelos com vetores de 300 dimensões, pré-treinados para Português, obtidos a partir do repositório do NILC (Hartmann et al., 2017) (CBOW e GloVe) e do fastText[8] (FastText).
- Métodos baseados em *sentence embeddings* obtidos a partir de modelos de linguagem neuronais BERT (Devlin et al., 2019). Mais pro-

[6] https://whoosh.readthedocs.io/

[7] Term Frequency - Inverse Document Frequency

[8] https://fasttext.cc/

priamente, cada pergunta foi codificada com recurso a um modelo BERT, carregado e disponibilizado localmente a partir da plataforma bert-as-a-service[9], com todos os parâmetros por defeito, à excepção do parâmetro *maximum length of sequences*, usado com o valor NONE para que o tamanho máximo das sequências fosse igual ao tamanho da pergunta mais longa no AIA-BDE. Foram testados dois modelos BERT pré-treinados, nomeadamente: BERT-Base, Multilingual Cased (Multi-BERT), treinado e disponibilizado pelos criadores do BERT[10] para 104 línguas, que permite representar o texto em vetores de dimensão 768; BERTimbau-large-portuguese-cased (Souza et al., 2020), especificamente para português, que codifica o texto em vetores de 1.024 dimensões.

Para cada um dos métodos anteriores, e para cada tipo de variação, a Tabela 5 mostra a proporção de variações corretamente associadas às perguntas originais. Confirma-se uma dificuldade variável, dependendo do tipo de variação, o que suporta a nossa opção por identificar esse tipo. Como seria de esperar, é mais fácil associar as variações geradas automaticamente (VG1 e VG2) à pergunta original, o que fica claro com a observação de que todos os métodos têm um melhor desempenho nessas duas variações. Relembramos que estas variações são geradas com recurso ao Google Translate e é normal apresentarem poucas alterações relativamente à forma das perguntas originais. Assim, não é de admirar que vários métodos apresentem uma taxa de acerto superior a 85% para a primeira resposta, e 90% considerando a presença no top-5. Ainda que inferior, o melhor desempenho nas variações VIN é claramente superior ao desempenho nas VMT e, principalmente, nas VUC. Isto sugere que as variações VIN são mais fáceis de associar automaticamente à pergunta original. Por outro lado, as variações VUC parecem ser as mais desafiantes, o que estará relacionado com o seu maior nível de criatividade, já referido na Secção 3.2.

Também não é fácil de identificar o melhor método para esta tarefa, já que este varia para diferentes tipos de variação. Por exemplo, nas variações geradas automaticamente, o BERT multilingue tem um dos melhores desempenhos, apenas batido pelo word2vec-CBOW no top-3 e top-5 das variações VG1. No entanto, nas restantes variações há vários métodos com um desempenho superior. Nas variações VIN, o BERTimbau tem o melhor desempenho (83.5% para a primeira), mas não muito superior ao word2vec-CBOW (82.2%) ou a um método mais simples, como o Whoosh+ (82.3%). Nas VUC, apesar do desempenho médio ser inferior, os três melhores métodos são os mesmos que para as VIN, ainda que numa ordem diferente, nomeadamente Whoosh+ (62.5%), BERTimbau (60.4%), word2vec-CBOW (59.1%). Finalmente, para as VMT, o cenáro é um pouco diferente, com o word2vec-CBOW a ser claramente o melhor (70.8%), o que mostra que este talvez seja o método mais equilibrado, já que o desempenho do BERTimbau nestas variações foi bastante inferior (47.6%).

Por se tratar de um modelo recente, responsável por avanços significativos em várias tarefas do Processamento de Linguagem Natural, seria expectável que o melhor desempenho fosse alcançado pelos modelos BERT, o que não acontece. Contudo, chamamos a atenção de que estes modelos são mais complexos que os restantes e podem fornecer diferentes representações para o texto, considerando, por exemplo, a representação em diferentes camadas ou a sua combinação. Para além de não termos explorado todas essas possibilidades, utilizamos as versões pré-treinadas destes modelos, que não afinamos (em inglês, *fine-tuned*) para a tarefa em questão. Apesar de não termos uma quantidade suficiente de dados do domínio alvo para este fim, uma possibilidade seria afinar os modelos para calcular a similaridade semântica textual em português, tal como alguns investigadores já fizeram (Rodrigues et al., 2020b) (Rodrigues et al., 2020a).

5. Classificação da Origem das Variações

Como referido na Secção 3.1, as perguntas do AIA-BDE foram obtidas a partir de quatro fontes principais: Espaço Empresa (EE), Regime Jurídico de Acesso e Exercício de Atividades de Comércio, Serviços e Restauração (RJACSR), Alojamento Local (AL) e Apoios Sociais (AS). A origem das perguntas pode ser vista como o seu assunto de alto nível, isto é, todas as perguntas com a mesma origem estarão relacionadas e serão acerca do mesmo tipo de serviços. Só por si, esta informação pode ser importante porque, em alguns casos, um agente de pesquisa ou resposta a perguntas pode começar por identificar o assunto ou domínio da pergunta do utilizador, e assim diminuir o conjunto onde procurar a resposta. No limite, tal agente poderá nem encontrar uma resposta adequada, mas pelo menos

[9] https://github.com/hanxiao/bert-as-service
[10] https://github.com/google-research/bert

Método	Variações														
	VG1 (855)			VG2 (855)			VIN (2,279)			VUC (816)			VMT (168)		
	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
Whoosh base	83.2	90.9	93.1	80.2	88.3	90.5	73.8	85.7	88.2	50.9	65.6	69.5	59.5	73.8	79.2
Whoosh+	88.1	95.6	96.6	85.4	93.7	95.6	82.3	91.4	94.0	**62.5**	**78.9**	**83.0**	54.8	70.2	81.5
CBOW	88.4	**95.8**	**97.1**	86.5	94.7	96.4	82.2	91.0	93.3	59.1	75.5	79.8	**70.8**	**86.3**	**89.9**
FastText	57.2	68.5	72.6	51.5	63.4	68.0	40.0	51.0	55.2	21.0	28.8	34.2	39.9	57.7	64.9
GloVe	85.1	92.0	93.9	82.5	90.3	92.2	70.0	79.2	81.7	46.3	58.0	64.6	63.7	79.8	83.3
Multi-BERT	**90.6**	95.3	96.6	**90.6**	**96.0**	**97.5**	73.7	84.0	87.1	46.4	59.6	65.8	39.9	53.0	56.5
BERTimbau	86.1	94.9	96.1	83.6	93.1	94.5	**83.5**	**92.4**	**94.1**	60.4	75.2	79.4	47.6	57.1	62.5

Tabela 5: Proporção de variações de diferentes tipos corretamente mapeadas com as perguntas originais (Top1), nas Top-3 e nas Top-5 mais similares, utilizando diferentes representações vetoriais.

conseguir indicar ao utilizador uma página web ou uma lista de perguntas sobre o assunto identificado.

De forma a simular este processo, outra experiência realizada com o AIA-BDE procurou, de forma automática, identificar a fonte das variações. Para tal, adoptamos uma abordagem de Aprendizagem Automática Supervisionada (em inglês, *Supervised Machine Learning*), em que classificadores foram treinados com as perguntas originais, e testados na identificação da origem das variações. Entre os classificadores testados, destacou-se um classificador baseado numa *Support Vector Machine* (SVM) linear, que utilizamos com duas representações diferentes do texto:

- Vetores TF-IDF com dimensão máxima 750 (i.e., comparável com a dos vetores do BERT-base), utilizando como *features* os átomos que ocorriam em pelo menos duas perguntas e, de forma a não considerar palavras demasiado frequentes, no máximo, em 50% de todas as perguntas originais.
- Vetores de dimensão 768, resultantes da codificação pelo modelo pré-treinado BERTimbau-base (Souza et al., 2020)[11], desta vez recorrendo à biblioteca `transformers`, da HuggingFace[12], e à pipeline *feature-extraction*.

A experiência foi realizada com recurso à biblioteca Python scikit-learn (Pedregosa et al., 2011) e respetivas implementações dos classificadores (i.e., LinearSVC com os parâmetros por omissão, para o classificador baseado em SVM), vetorização TF-IDF (TfidfVectorizer), e cálculo de métricas. Ao optar pelo BERT, os vetores deste último eram utilizados em alternativa aos vetores TF-IDF.

As Tabelas 6 e 7 apresentam o desempenho nesta experiência usando os vetores TF-IDF e BERT. Em cada uma, incluímos a precisão (P), abrangência (A) e medida-F1 (F1) do modelo treinado com todas as perguntas e testado com as variações de cada tipo. Estas métricas são apresentadas para cada classe (i.e., origem) e também para o total, através de uma macro-média (Macro-μ), onde a proporção de instâncias de cada classe não é considerada, e de uma média pesada (μ-pesada). Para as variações que não cobrem alguma classe, as classes em falta não foram consideradas no cálculo das médias.

Uma vez mais, verifica-se que, dependendo da variação, pode ser mais ou menos difícil identificar a origem. Com a representação mais tradicional, TF-IDF, as variações onde o classificador teve mais dificuldades foram as VUC, o que volta a sugerir que são aquelas que mais se desviam das perguntas originais. Foi também para estas variações que o desempenho foi claramente superior com as representações baseadas no BERT, o que mostra a capacidade deste modelo lidar com diferenças lexicais em textos com o mesmo significado. Ainda que o desempenho para as variações VMT seja superior ao desempenho para as VUC, o melhor desempenho para ambas é obtido com os vetores baseados no BERT. Como referido anteriormente, e passando a redundância, estas são nada mais nada menos do que as variações onde há maior variação lexical relativamente à pergunta original. Por outro lado, para os outros três tipos de variação, a utilização do BERT parece não trazer grandes benefícios, apresentando mesmo um desempenho ligeiramente inferior.

Importa ainda destacar o desempenho superior na classificação de variações VIN, que, utilizando os vetores TF-IDF, atinge uma medida-F1 próxima do 100%, quando para as VG1 e VG2 este valor se situa em torno dos 90%. Acreditamos que isto também seja uma consequência do método adoptado para a geração automática de variações. Mais propriamente, a uma pequena proporção de resultados que, devido a problemas de tradução, nomeadamente do nome de serviços, introduzem termos fora do esperado e, consequentemente, fora do domínio de cada classe.

[11] Neste caso optámos pela versão do modelo *base*, mais simples, porque, com a versão *large*, a SVM tinha dificuldade em convergir, e por isso a melhorar os resultados.

[12] `https://huggingface.co/transformers/`

Origem	**VG1**			**VG2**			**VUC**			**VIN**			**VMT**		
	P	A	F1	P	A	F1	P	A	F1	P	A	F1	P	A	F1
EE	91%	99%	95%	91%	99%	95%	65%	99%	78%	100%	98%	99%	N/A	N/A	N/A
RJACSR	92%	67%	77%	90%	69%	78%	96%	49%	65%	N/A	N/A	N/A	N/A	N/A	N/A
AL	93%	75%	83%	98%	71%	82%	99%	69%	81%	N/A	N/A	N/A	N/A	N/A	N/A
AS	100%	86%	92%	100%	84%	91%	N/A	N/A	N/A	N/A	N/A	N/A	100%	70%	83%
Macro-μ	**94%**	82%	**87%**	**95%**	**81%**	**87%**	87%	72%	75%	**100%**	**98%**	**99%**	**100%**	70%	83%
μ-pesada	**92%**	**92%**	**92%**	**92%**	**92%**	**92%**	82%	74%	73%	**100%**	**98%**	**99%**	**100%**	70%	83%

Tabela 6: Classificação da origem das variações, com SVM e representação TF-IDF.

Origem	**VG1**			**VG2**			**VUC**			**VIN**			**VMT**		
	P	A	F1	P	A	F1	P	A	F1	P	A	F1	P	A	F1
EE	93%	92%	93%	92%	95%	94%	87%	93%	89%	100%	91%	95%	N/A	N/A	N/A
RJACSR	61%	68%	64%	70%	63%	66%	87%	85%	86%	N/A	N/A	N/A	N/A	N/A	N/A
AL	91%	95%	93%	91%	89%	90%	86%	66%	74%	N/A	N/A	N/A	N/A	N/A	N/A
AS	96%	89%	93%	89%	89%	89%	N/A	N/A	N/A	N/A	N/A	N/A	100%	81%	89%
Macro-μ	85%	**86%**	86%	87%	**81%**	83%	**90%**	**86%**	**87%**	**100%**	91%	95%	**100%**	**81%**	**89%**
μ-pesada	89%	88%	89%	89%	89%	89%	**87%**	**86%**	**86%**	**100%**	91%	95%	**100%**	**81%**	**89%**

Tabela 7: Classificação da origem das variações, com SVM e representação BERT.

Vejam-se alguns exemplos onde isto acontece:

P: *Há sociedades que não podem ser constituídas nos balcões "Empresa na Hora"?*

VG1: *Existem empresas que não podem ser configuradas nos contadores "Empresa no Horário"?*

VG2: *Existem empresas que não podem ser configuradas nos contadores "Business on Time"?*

P: *A que balcão de atendimento "Empresa na Hora" me devo dirigir?*

VG1: *Qual service desk devo entrar em contato?*

Apesar do AIA-BDE sofrer de desequilíbrio ao nível da origem, o desempenho individual em cada origem não parece ser muito afetado pelo número de perguntas originais com essa origem. É de notar, aliás, que o desempenho inferior é claramente para as variações RJACSR, quando o número de perguntas com esta origem (118) é mais do dobro das perguntas de AL e AS (56). Uma possível explicação seria o facto de estas perguntas serem mais longas que as demais, mas como a Tabela 1 mostra, as perguntas de AL são as mais longas, com uma diferença considerável. Uma última possibilidade será um maior vocabulário utilizado por estas perguntas. No entanto, identificar claramente a razão para este desempenho inferior implicaria uma análise mais profunda das perguntas.

Após esta experiência, não fica claro se seria benéfica a utilização de um classificador inicial que fizesse a triagem das perguntas de acordo com a sua origem. Com a excepção das variações VIN, haveria ainda uma proporção considerável de perguntas mal classificadas, o que impossibilitaria à partida a identificação da sua resposta correta. No entanto, tal como acontece para a experiência anterior, estes resultados devem ser vistos como ilustrativos daquilo que pode ser feito com o AIA-BDE, e apenas uma base (*baseline*) com grande margem de melhoria. Por exemplo, à semelhança da experiência anterior, poderiam ter sido obtidas representações alternativas a partir do modelo BERT, ou utilizada uma versão deste modelo afinada (*fine-tuned*) para a classificação automática, sem recurso ao classificador SVM.

6. Classificação do Tipo de Pergunta

Considerando que o tipo de uma pergunta pode condicionar a resposta a dar e o processo de a obter, numa última experiência procuramos identificar automaticamente esse tipo. Contudo, aqui deparámo-nos com dois problemas, também referidos na Secção 3.3: (i) as anotações do tipo de pergunta foram muito variáveis, com apenas cerca de um quarto das perguntas em que os cinco anotadores concordaram; (ii) considerando estas anotações, o corpo AIA-BDE é altamente desequilibrado relativamente ao tipo de pergunta com maior confiança.

Por se tratar apenas de uma experiência ilustrativa do que é possível fazer com o AIA-BDE, optamos por realizar algumas simplificações. Mais propriamente, consideramos que cada pergunta só podia ter um tipo, e que esse seria o tipo em que tínhamos maior confiança. Isto eliminou automaticamente da nossa experiência as per-

guntas em que havia tipos empatados, com 40% ou mesmo 20% de confiança. Para lidar com o segundo problema, decidimos focar-nos apenas nos tipos para os quais, após a aplicação da condição anterior, restava um número aceitável de exemplos.

A Tabela 8 apresenta os quatro tipos com mais de 30 perguntas (Definição, Procedimento, Tempo, Condição), e a respetiva quantidade de perguntas desse tipo, após a simplificação anterior. Ao verificarmos que para dois dos quatro tipos anteriores, Tempo e Condição, há pouco mais de 40 perguntas, decidimos realizar a experiência de duas formas: uma considerando os quatro tipos, outra considerando apenas os dois maioritários, Definição e Procedimento.

Tipo	Quantidade
Definição	204
Procedimento	132
Tempo	44
Condição	43

Tabela 8: Distribuição após remoção de instâncias com menor confiança.

A experiência inicial consistiu em treinar um classificador para prever o tipo da pergunta, dada o texto da pergunta. No entanto, acabamos também por experimentar até que ponto seria possível fazer o mesmo, mas considerando apenas o texto da resposta.

Voltamos a experimentar um conjunto de classificadores, incluídos na biblioteca scikit-learn, e baseados nas mesmas representações vetoriais da experiência anterior (Secção 5), TF-IDF e BERT. Uma vez mais, o classificador baseado numa SVM linear voltou a destacar-se. Foi também utilizado com os parâmetros por omissão, desta vez com a excepção do número máximo de iterações. Ao verificarmos, através de uma mensagem de *warning*, que nem sempre havia convergência, decidimos aumentar o número máximo de iterações de 1.000, o valor por omissão, para 3.000, minimizando desta forma o problema.

Outra diferença desta experiência foi a existência de menos dados e inexistência de uma separação clara em dados a usar para treino e para teste. Assim, nas Tabelas 9 e 10, optamos por apresentar os resultados de uma validação cruzada em 10 subconjuntos (em inglês, *10-fold cross validation*), respetivamente para as experiências com quatro e dois tipos. Os resultados podem ser analisados com base na exatidão (em inglês, *accuracy*), i.e., a proporção de tipos corretamente identificados, mas também através das macro médias da precisão (Macro-P), abrangência (Macro-A) e medida-F1 (Macro-F1). Enquanto que na *accuracy* os tipos mais frequentes terão um maior peso para a média, nas macro médias o desempenho em cada tipo tem o mesmo peso.

Os desempenhos reportados mostram que, mesmo com as simplificações realizadas, a identificação do tipo de pergunta é desafiante. Esta situação é mais evidente quando se consideram quatro tipos, e ainda mais quando a identificação se baseia na resposta e não na pergunta. O impacto do desequilíbrio observa-se nos valores da medida-F1, sempre mais baixos do que a *accuracy*, principalmente quando são considerados quatro tipos. Ou seja, como seria de esperar, o desempenho será melhor para os tipos mais representados.

De notar ainda os desvios padrão elevados, que mostram que o desempenho depende muito da escolha do conjunto de treino, mesmo quando esse conjunto inclui 90% dos dados (i.e., validação cruzada em 10 subconjuntos). Estes desvios dificultam a análise da melhor forma de representação e não é possível tirar grandes conclusões. Mesmo que, por exemplo, no cenário com quatro tipos, as médias sugiram que a representação BERT funcione melhor quando se usa a pergunta, e que a representação TF-IDF seja preferível quando se usa a resposta, os desvios padrão, respetivamente 10% e 8%, mostram que também pode acontecer o contrário.

7. Conclusões

Apresentamos neste artigo o corpo AIA-BDE, focado em FAQ que cobrem um pequeno conjunto de domínios da Administração Pública de Portugal. Apesar de outros ficheiros complementares, focamo-nos em 855 perguntas, para as quais variações foram produzidas manual e automaticamente, permitindo assim a avaliação de sistemas de recuperação de FAQ, que podem estar integrados em sistemas de resposta automática a perguntas, ou até de diálogo. Para além das variações, a cada uma das 855 perguntas anteriores está associada uma lista de tipos de pergunta e respetiva confiança.

Utilizamos ainda o AIA-BDE em três experiências, com vista à avaliação de: (i) utilização das variações como perguntas de um utilizador, e respetiva associação a perguntas conhecidas; (ii) classificação de variações de acordo com a origem da respetiva pergunta original; (iii) identificação do tipo de pergunta. As experiências realizadas ajudaram-nos a compreender melhor

Entrada	Modelo	*Accuracy*	Macro-P	Macro-A	Macro-F1
Pergunta	SVM + TFIDF	86±5%	85±10%	78±7%	79±8%
	SVM + BERT	87±8%	84±10%	83±11%	82±10%
Resposta	SVM + TFIDF	73±6%	66±12%	60±9%	61±10%
	SVM + BERT	69±9%	61±14%	58±11%	56±11%

Tabela 9: Validação cruzada para a identificação do tipo de pergunta para as quatro classes com mais de 40 instâncias: Definição, Procedimento, Condição, Tempo

Entrada	**Modelo**	*Accuracy*	Macro-P	Macro-A	Macro-F1
Pergunta	SVM + TFIDF	93±2%	92±2%	92±3%	92±3%
	SVM + BERT	94±4%	94±4%	93±5%	93±4%
Resposta	SVM + TFIDF	82±7%	82±8%	80±7%	80±7%
	SVM + BERT	82±7%	84±7%	81±7%	81±7%

Tabela 10: Validação cruzada da identificação do tipo de pergunta para as duas classes com mais de 100 instâncias: Definição e Procedimento

o seu conteúdo, e demonstraram a sua utilidade, ao mesmo tempo que estabeleceram resultados base (*baselines*), com margem para melhoria no futuro. Verificamos que diferentes tipos de variação trazem desafios diferentes e que, regra geral, não parece haver um método que se adapte bem a todos os tipos. O mesmo acontece com as perguntas de diferentes origens. Por exemplo, verificamos que, apesar das representações obtidas a partir de modelos de linguagem baseados em *transformers* (BERT) levarem a resultados interessantes, isso não se verifica para todo o tipo de variação, e uma abordagem baseada em IR tradicional, bem mais simples, pode ser bastante competitiva. Ainda assim, acreditamos que as abordagens baseadas em *transformers* tenham maior margem de progresso, por exemplo, se forem afinadas para o cálculo da similaridade semântica; ou se forem pré-treinados em dados específico do domínio. O mesmo para a abordagens baseadas em *word embeddings*, onde poderia adicionalmente ser benéfico considerar expressões ou entidades multipalavra (e.g., *Cartão Provisório de Identificação de Pessoa Colectiva* ou *Empresa na Hora*).

Depois da terceira experiência, verificamos também que seria importante rever os tipos de pergunta atualmente considerados, tentando diminuir ou uniformizar as sobreposições possíveis. Paralelamente, poderia ser interessante abordar a classificação automática do tipo como um problema de multi-classificação.

O AIA-BDE é disponibilizado à comunidade através de `https://github.com/NLP-CISUC/AIA-BDE`, para que possa ser utilizado em experiências como aquelas aqui apresentadas, com vista à melhoria dos resultados base, ou em experiências alternativas. Apesar de focado num domínio, acreditamos que as variações podem ser usadas como dados de treino para a identificação mais generalizada de paráfrases em contexto interrogativo, ou como base para a criação de uma coleção para a avaliação da similaridade semântica textual no mesmo contexto. O AIA-BDE pode ainda servir de base a um agente que responde a perguntas acerca dos domínios cobertos. Aqui, as variações podem servir apenas para avaliação do sistema, como feito recentemente (Santos et al., 2020a), mas também usadas como variação das intenções no processo de compreensão de linguagem natural (em inglês, *Natural Language Understanding*).

Ainda que não seja uma prioridade, no futuro poderão ser incluídas mais perguntas e mais variações ao corpo, aumentando assim a sua dimensão e tornando-o mais apto para o treino de modelos mais poderosos. Para além do AIA-BDE, mostramos que um sistema de recuperação de FAQ pode ser avaliado com base num conjunto de variações para cada pergunta, que simulem as necessidades de informação dos utilizadores. Estas necessidades podem, em alguns casos, ser expressas de uma forma semelhante àquela em que as perguntas estão armazenadas mas, devido à variabilidade linguística, podem também ser colocadas de formas radicalmente diferentes, ao nível lexical ou sintático. Apesar do trabalho manual envolvido, acreditamos que é uma vantagem ter um recurso deste tipo, que permita ir avaliando progressos desta forma. O processo de criação poderá ser replicado para outros domínios e, ainda que possa

ser benéfico recorrer a especialistas do domínio para o fazer, o mais importante é cobrir diferentes necessidades de informação, incluindo aquelas de utilizadores menos experientes, obtidas, por exemplo, com recurso a *crowdsourcing*.

Agradecimentos

Parte deste trabalho foi realizado no âmbito do projeto demonstrador AIA, "Apoio Inteligente a empreendedores (chatbots)", financiado pela FCT, através da iniciativa INCoDe 2030.

Gostaríamos também de agradecer: ao João Ferreira, pelo seu envolvimento na integração das variações mais recentes no corpo e na definição dos tipos de pergunta; à Luísa Coheur e aos seus alunos, pela criação das variações VIN; à AMA, em especial ao Jorge Cabrita de Sousa, por nos ceder uma grande parte dos materiais de onde foram extraídas as perguntas originais.

Referências

Agirre, Eneko, Mona Diab, Daniel Cer & Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. Em *1st Joint Conference on Lexical and Computational Semantics: 6th International Workshop on Semantic Evaluation*, 385–393.

Bowman, Samuel R., Gabor Angeli, Christopher Potts & Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. Em *Conference on Empirical Methods in Natural Language Processing*, 632–642. 10.18653/v1/D15-1075.

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*, 5016–5026. 10.18653/v1/D18-1547.

Burke, Robin D, Kristian J Hammond, Vladimir Kulyukin, Steven L Lytinen, Noriko Tomuro & Scott Schoenberg. 1997. Question answering from frequently asked question files: Experiences with the FAQ finder system. *AI magazine* 18(2). 57–57.

Caputo, Annalina, Marco Degemmis, Pasquale Lops, Francesco Lovecchio & Vito Manzari. 2016. Overview of the EVALITA 2016 question answering for frequently asked questions (QA4FAQ) task. Em *3rd Italian Conference on Computational Linguistics (CLiC-it): 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA)*, CEUR–WS.

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

Choi, Eunsol, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang & Luke Zettlemoyer. 2018. QuAC: Question answering in context. Em *Conference on Empirical Methods in Natural Language Processing*, 2174–2184. 10.18653/v1/D18-1241.

Criscuolo, Marcelo, Erick Rocha Fonseca, Sandra Maria Aluísio & Ana Carolina Sperança-Criscuolo. 2017. MilkQA: a dataset of consumer questions for the task of answer selection. Em *6th Brazilian Conference on Intelligent Systems (BRACIS)*, 354–359. 10.1109/BRACIS.2017.12.

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*, 4171–4186. 10.18653/v1/N19-1423.

Fellbaum, Christiane (ed.). 1998. *WordNet: An Electronic Lexical Database (language, speech, and communication)*. The MIT Press.

Fernandes, Mariana Gaspar, Cátia Dias & Luísa Coheur. 2019. Distinguishing different classes of utterances - the UC-PT corpus. Em *8th Symposium on Languages, Applications and Technologies (SLATE)*, 14:1–14:8. 10.4230/OASIcs.SLATE.2019.14.

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

Forner, Pamela, Anselmo Peñas, Eneko Agirre, Iñaki Alegria, Corina Forăscu, Nicolas Moreau, Petya Osenova, Prokopis Prokopidis, Paulo Rocha, Bogdan Sacaleanu et al. 2008. Overview of the CLEF 2008 multilingual Question Answering track. Em *Workshop of the Cross-Language Evaluation Forum for European Languages*, 262–295.

Gonçalo Oliveira, Hugo, Ricardo Filipe, Ricardo Rodrigues & Ana Alves. 2019. Using Lucene for developing question-answering agent in Portuguese. Em *8th Symposium on Languages, Applications and Technologies (SLATE)*, 2:1–2:14. 10.4230/OASIcs.SLATE.2019.2.

Guillou, Pierre. 2021. Portuguese BERT base cased QA (Question Answering), finetuned on SQUAD v1.1. https://huggingface.co/pierreguillou/bert-base-cased-squad-v1.1-portuguese.

Gunasekara, Chulaka, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, Dilek Hakkani-Tür, Jinchao Li, Qi Zhu, Lingxiao Luo, Lars Liden, Kaili Huang, Shahin Shayandeh, Runze Liang, Baolin Peng, Zheng Zhang, Swadheen Shukla, Minlie Huang, Jianfeng Gao, Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, Maxine Eskenazi, Ahmad Beirami, Eunjoon, Cho, Paul A. Crook, Ankita De, Alborz Geramifard, Satwik Kottur, Seungwhan Moon, Shivani Poddar & Rajen Subba. 2020. Overview of the 9th dialog system technology challenge: DSTC9. ArXiv:2011.06486 [cs.CL].

Hartmann, Nathan S., Erick R. Fonseca, Christopher D. Shulby, Marcos V. Treviso, Jéssica S. Rodrigues & Sandra M. Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. Em *11th Brazilian Symposium in Information and Human Language Technology (STIL)*, 122–131.

Karan, Mladen & Jan Šnajder. 2016. FAQIR–a frequently asked questions retrieval test collection. Em *International Conference on Text, Speech, and Dialogue*, 74–81.

Karan, Mladen & Jan Šnajder. 2018. Paraphrase-focused learning to rank for domain-specific frequently asked questions retrieval. *Expert Systems with Applications* 91. 418–433. 10.1016/j.eswa.2017.09.031.

Karan, Mladen, Lovro Žmak & Jan Šnajder. 2013. Frequently asked questions retrieval for Croatian based on semantic textual similarity. Em *4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, 24–33.

Kolomiyets, Oleksandr & Marie-Francine Moens. 2011. A survey on question answering technology from an information retrieval perspective. *Information Sciences* 181(24). 5412–5434. 10.1016/j.ins.2011.07.047.

Lison, Pierre & Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. Em *10th International Conference on Language Resources and Evaluation (LREC)*, 923–929.

Lowe, Ryan, Nissan Pow, Iulian Serban & Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. Em *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 285–294. 10.18653/v1/W15-4640.

Magarreiro, Daniel, Luísa Coheur & Francisco S. Melour. 2014. Using subtitles to deal with out-of-domain interactions. Em *18th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*, 98–106.

Magnini, Bernardo, Alessandro Vallin, Christelle Ayache, Gregor Erbach, Anselmo Peñas, Maarten de Rijke, Paulo Rocha, Kiril Ivanov Simov & Richard F. E. Sutcliffe. 2004. Overview of the CLEF 2004 Multilingual Question Answering track. Em *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum (CLEF), Revised Selected Papers*, 371–391.

Mota, Cristina, Alberto Simões, Cláudia Freitas, Luís Costa & Diana Santos. 2012. Págico: Evaluating Wikipedia-based information retrieval in Portuguese. Em *8th International Conference on Language Resources and Evaluation (LREC)*, 2015–2022.

Nakov, Preslav, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin & Karin Verspoor. 2017. SemEval-2017 task 3: Community question answering. Em *11th International Workshop on Semantic Evaluation (SemEval)*, 27–48. 10.18653/v1/S17-2003.

Nakov, Preslav, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass & Bilal Randeree. 2015. SemEval-2015 task 3: Answer selection in community question answering. Em *9th International Workshop on Semantic Evaluation (SemEval)*, 269–281. 10.18653/v1/S15-2047.

Nakov, Preslav, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass & Bilal Randeree. 2016. SemEval-2016 task 3: Community question answering. Em *10th International Workshop on Semantic Evaluation*, 525–545. 10.18653/v1/S16-1083.

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

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev & Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. Em *Conference on Empirical Methods in Natural Language Processing*, 2383–2392.

Real, Livy, Erick Fonseca & Hugo Gonçalo Oliveira. 2020. The ASSIN 2 shared task: a quick overview. Em *13th International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 406–412. 10.1007/978-3-030-41505-1_39.

Reddy, Siva, Danqi Chen & Christopher D Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics* 7. 249–266.

Ritter, Alan, Colin Cherry & William B Dolan. 2011. Data-driven response generation in social media. Em *Conference on empirical methods in natural language processing*, 583–593.

Rodrigues, Ruan Chaves, Jéssica Rodrigues da Silva, Pedro Vitor Quinta de Castro, Nádia Félix Felipe da Silva & Anderson da Silva Soares. 2020a. Multilingual transformer ensembles for Portuguese natural language tasks. Em *ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*, CEUR–WS.

Rodrigues, Rui, Paula Couto & Irene Rodrigues. 2020b. IPR: The semantic textual similarity and recognizing textual entailment systems. Em *ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*, CEUR–WS.

Santos, Diana & Paulo Rocha. 2004. The key to the first CLEF with Portuguese: topics, questions and answers in CHAVE. Em *Workshop of the Cross-Language Evaluation Forum for European Languages*, 821–832.

Santos, José, Luís Duarte, João Ferreira, Ana Alves & Hugo Gonçalo Oliveira. 2020a. Developing Amaia: A conversational agent for helping Portuguese entrepreneurs – An extensive exploration of question-matching approaches for Portuguese. *Information* 11(9). 10.3390/info11090428.

Santos, José, Ana Alves & Hugo Gonçalo Oliveira. 2020b. Leveraging on Semantic Textual Similarity for developing a Portuguese dialogue system. Em *13th International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 131–142. 10.1007/978-3-030-41505-1_13.

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

Vinyals, Oriol & Quoc V. Le. 2015. A neural conversational model. Em *International Conference on Machine Learning, Deep Learning Workshop*, arXiv:1506.05869 [cs.CL].

Voorhees, Ellen M. 2008. Evaluating question answering system performance. Em *Advances in Open Domain Question Answering*, 409–430. 10.1007/978-1-4020-4746-6_13.

Wen, Tsung-Hsien, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes & Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. Em *15th Conference of the European Chapter of the Association for Computational Linguistics*, 438–449.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q Weinberger & Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. Em *8th International Conference on Learning Representations (ICLR)*, arXiv:1904.09675 [cs.CL].