Instituto Superior
de Engenharia
Politécnico de Coimbra

DEPARTAMENTO DE INFORMÁTICA E SISTEMAS

Descoberta de Fluxos de Diálogo através de Técnicas de Aprendizagem Computacional

Trabalho de Projeto para a obtenção do grau de Mestre em Engenharia Informática

Especialização em Análise Inteligente de Dados

Autor

Daniel Lemos Martins

Orientadores

Professora Doutora Ana Alves

Professor Doutor Hugo Gonçalo Oliveira

INSTITUTO POLITÉCNICO
DE COIMBRA

INSTITUTO SUPERIOR
DE ENGENHARIA
DE COIMBRA

Coimbra, março 2023

RESUMO

Os centros de atendimento são alvo de elevada pressão para atingir altos níveis de exigência, fruto da migração de mercados para o mundo digital. Os utilizadores cada vez mais esperam ter os seus problemas resolvidos em tempo real e à distância de um clique. Por esta razão, estes serviços tornam-se essenciais na relação cliente-empresa. O sucesso ou insucesso da interação entre um cliente e um centro de atendimento depende principalmente dos agentes envolvidos, sejam humanos ou virtuais, pois são eles que desempenham o papel principal na comunicação com o utilizador.

Com os avanços da Inteligência Artificial e de *Machine Learning* tem-se procurado desenvolver soluções capazes de automatizar ou acelerar a comunicação com o utilizador. No entanto, são muitos os desafios encontrados para atingir o objetivo final: uma otimização dos agentes inteligentes para que ofereçam uma resposta rápida e eficaz. Desta forma, é imperativo manter completo e atualizado o conhecimento destes agentes para que possam responder às necessidades e requisitos dos clientes.

Neste sentido, o presente trabalho utiliza técnicas de Processamento de Linguagem Natural e recorre à aprendizagem não supervisionada para agrupar falas em diferentes conjuntos de diálogos públicos de acordo com o seu significado. São utilizadas diferentes representações das falas, que levam a uma análise sobre o que os *clusters* obtidos representam e do que se aproximam. Esse *clustering* permite identificar sequências entre conjuntos relacionados de falas, e assim representar fluxos de diálogo. Estes permitem dar uma visão de alto nível de conversas, que servirão de auxílio aos agentes inteligentes e, consequentemente, aos centros de atendimento. Esse auxílio foi testado através da criação de um *chatbot* que tenha como orientação os fluxos que foram produzidos.

Através de abordagens não supervisionadas propõe-se uma metodologia para criar fluxos de diálogo que podem ser uma ferramenta valiosa para melhorar a eficiência de sistemas de recomendação.

Palavras-chave: Processamento de Linguagem Natural; Reconhecimento de Atos de Diálogo; Reconhecimento de *Intents*; *Machine Learning*; Modelos de Markov; *Chatbots*.

ABSTRACT

Call centers are under immense pressure to meet very high standards, as markets migrate to the digital world. Users increasingly expect to have their problems solved in real time and at the distance of a click. For this reason, this service becomes essential in the customer-company relationship. The success or failure of the interaction between a customer and a call center depends mainly on the agents involved, whether human or digital, as they play the main role in communication with the user.

With the advances in Artificial Intelligence and Machine Learning, there has been an attempt to develop solutions capable of automating or accelerating communication with the user. However, there are many challenges to achieve the final goal: an optimization of intelligent agents to provide a fast and effective response. Thus, it is imperative to keep the knowledge of these agents complete and up to date so that they can respond to the customer needs and requirements.

In that regard, the present work uses Natural Language Processing techniques and resorts to unsupervised learning to cluster dialogues from different public conversations according to their meaning. Different representations of the dialogues are used, which lead to an analysis of what the obtained clusters represent and what they come close to. This clustering allows for the identification of sequences between related sets of dialogues, thus representing dialogue flows. These provide a high-level view of conversations, which will serve as an aid to intelligent agents and, consequently, call centers as they will have an accurate control throughout a conversation. This assistance was tested through the creation of a chatbot that was guided by the flows that were produced.

Through unsupervised approaches, a methodology is proposed to create dialogue flows that can be a valuable tool for improving the efficiency of recommendation systems.

Keywords: Natural Language Processing; Dialog Act Discovery; Intent Discovery; Machine Learning; Markov Models; Chatbots.

AGRADECIMENTOS

Este trabalho só foi possível pela contribuição de diversas pessoas que me ajudaram a ultrapassar as diversas barreiras que foram aparecendo neste longo e árduo caminho, que agora acabando, é reconfortante, e por isso o meu agradecimento é imprescindível.

Queria agradecer a todos os docentes que fizeram parte do meu percurso, desde o pré-ensino ao ensino superior, cada um à sua forma me marcou e contribuiu para o meu desenvolvimento enquanto pessoa.

Aos professores Doutores Ana Alves e Hugo Gonçalo Oliveira, os meus orientadores, que foram incansáveis e estiveram sempre disponíveis para me ajudar e guiar ao longo deste projeto. E que desde o início sempre me mostraram o caminho para alcançar os diversos objetivos do trabalho.

Aos meus colegas de projeto, nomeadamente à professora Doutora Catarina e à Patrícia, pelo apoio e ajuda dada nestes meses.

Aos meus amigos que estiveram presentes no meu percurso, sempre acreditaram em mim e me motivaram a continuar.

O meu trabalho de Mestrado foi apoiado financeiramente pela bolsa FLOWANCE com a referência 793050, co-financiado pelo FEDER no âmbito do Projeto FLOWANCE (POCI-01-0247-FEDER-047022).

Um agradecimento à minha família, mas em especial aos meus pais e ao meu irmão, que são as pessoas mais importantes na minha vida e estiveram sempre presentes nos bons e nos maus momentos.

Por último, quero agradecer a mim próprio, porque apesar de todas as dificuldades apresentadas ao longo deste caminho, fui resiliente e nunca desisti.

A todos, o meu sincero obrigado.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE SIGLAS, ACRÓNIMOS E SÍMBOLOS

CNN	*Convolutional Neural Network*
DAC	*Dialog Act Classification*
EOD	*End of Dialogue*
GRU	*Gated Recurrent Unit*
IA	Inteligência Artificial
ISO	*International Organization for Standardization*
LDA	*Latent Dirichlet Allocation*
LSTM	*Long Short-Term Memory*
ML	*Machine Learning*
NER	*Named Entity Recognition*
PCA	*Principal Component Analysis*
PLN	Processamento de Linguagem Natural
RNN	*Recurrent Neural Network*
SGD	*Schema Guided Dialogue*
SOD	*Start of Dialogue*
t-SNE	*t-distributed Stochastic Neighbor Embedding*
TF-IDF	*Term Frequency-Inverse Document Frequency*

1 INTRODUÇÃO

A expansão tecnológica tem possibilitado o surgimento de novas oportunidades e, consequentemente, novos desafios fruto da constante pressão e expectativas elevadas dos utilizadores. Um exemplo disso são os centros de atendimento para clientes, onde são exigidas respostas rápidas em tempo real que sejam concisas e capazes de resolver os diferentes problemas. A relação entre os clientes e as empresas é um fator que determina o sucesso do negócio, e por isso é imprescindível a satisfação dos clientes.

Com tanta oferta nos dias de hoje, a exigência dos clientes também aumenta, e uma simples falha pode fazer ruir um negócio. O papel dos centros de atendimento é crucial na sustentação da relação consumidor-empresa uma vez que permite responder a variadas questões e/ou necessidades que possam surgir aquando da utilização do serviço. Deste modo, é imperativo alcançar um atendimento robusto.

Desta forma, são utilizadas diversas abordagens de Inteligência Artificial (IA) e *Machine Learning* (ML) na busca de uma resposta mais eficiente para melhorar a qualidade do atendimento e atender às expectativas do público.

São muitos os desafios inerentes à criação de um *chatbot*, ou agente virtual: os formatos de mensagem, sendo áudio ou texto, a uniformização da representação dos dados, a análise do texto com recurso a técnicas de Processamento de Linguagem Natural (PLN), o conhecimento usado pelos agentes de forma a tornar autêntica uma determinada interação. Esta área estuda a comunicação entre humanos, incluindo a análise e geração de diálogos, sejam eles escritos ou falados, de forma a melhorar a interação homem-máquina e de compreender automaticamente textos. Portanto, um diálogo corresponde a um conjunto de falas, que podem ser classificadas quanto à sua natureza em atos de diálogo ou *intents*. Os atos de diálogo correspondem às ações num diálogo, questionar, saudar, agradecer, etc. Os *intents* regerem-se à intenção do utilizador, como agendar uma consulta, saber informações sobre produtos, etc.

São exploradas técnicas para agrupar falas de acordo com o seu significado /ação subjacente. Essa possibilidade de descobrir padrões nos dados de *clusters* semelhantes e o facto de não existir dados anotados para todos os domínios motiva a escolha de aprendizagem não supervisionada. Posteriormente é feita uma análise para verificar até que ponto esses *clusters* se aproximam de atos de diálogo ou *intents*. Este *clustering* possibilitará a classificação e identificação de sequências entre as diversas falas, criando os chamados fluxos de diálogo.

1.1 Motivações e Objetivos

Este projeto surgiu da necessidade da criação e do desenvolvimento de ferramentas que consigam criar sistemas de orientação/recomendação para facilitar o papel de atendimento dos clientes. Através dos fluxos de diálogo, é possível auxiliar agente humanos (ou mesmo agente inteligentes) com uma monitorização efetiva e atempada ao agente ao longo de uma conversa de modo a satisfazer as necessidades do cliente.

Deste modo, os objetivos principais são os de:

1. Extrair e agrupar em *clusters* diversas falas de diálogos e perceber que características dos dados permitem o *clustering*;
2. Comparação das diferentes formas de representação de falas, tendo em conta o *clustering* realizado;
3. Criação a grafos, de forma a ajudar a uma melhor interpretação dos fluxos de diálogo e tendências na comunicação.
4. Testar os grafos obtidos num *chatbot* e deste modo perceber como essa orientação pode ser útil para um agente humano ou artificial.

1.2 Contextualização

Este relatório tem como propósito documentar todo o trabalho desenvolvido ao longo desta dissertação, que está inserida no plano curricular do Mestrado em Engenharia Informática, ramo de Análise Inteligente de Dados, do Instituto Superior de Engenharia de Coimbra.

O trabalho realizado está enquadrado no projeto "FLOWANCE: *Automatic Dialogue Flow Extraction and Guidance*" (POCI-01-0247-FEDER-047022), co-financiado pelo Fundo Europeu de Desenvolvimento Regional (FEDER), através do programa Portugal 2020 (PT2020), e pelo Programa Operacional Competitividade e Internacionalização (COMPETE 2020). O projeto FLOWANCE tem como principal objetivo o desenvolvimento de soluções capazes de guiar a ação de um agente de um centro de contactos com base no histórico de interações entre agentes e clientes desse mesmo centro. O contributo deste projeto será a identificação de ferramentas capazes de representar fluxos de diálogo de forma intuitiva e que expressem quão provável são as transições entre tipos de falas, de modo a servir como guia orientador para o agente que irá ser criado. É possível observar na Figura 1.1 o esquema que representa a solução desejada na qual este trabalho está inserido. O foco deste trabalho é a componente de representação, em que diálogos já extraídos, de forma estruturada, terão de ser agrupados em *clusters*, de modo a permitir a criação de grafos que sirvam como base de orientação de um agente.

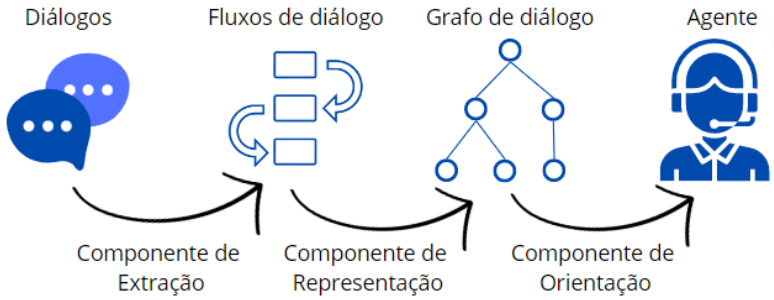


Figura 1.1 - Esquema da estrutura do projeto

1.3 Metodologias de Trabalho

Foram utilizadas diversas ferramentas de forma a agilizar o trabalho realizado e o seu contributo para este projeto. Nomeadamente utilizou-se o Slack[1] com um canal do projeto, que foi utilizado para tirar dúvidas pontuais, ou esclarecimento de alguma tarefa. Para a realização de reuniões, que durante quase todo o trabalho foram feitas semanalmente utilizou-se o Zoom[2] e Google Meet[3]. O Google Drive[4] foi utilizado para partilhar com os elementos do projeto os documentos que foram utilizados, nomeadamente ficheiros de código, os corpora utilizados. Utilizou-se o Google Colab[5] para a produção do código deste trabalho que foi desenvolvido com recurso à linguagem de programação Python[6].

1.4 Plano de Trabalho

Este projeto consistiu na execução das seguintes principais tarefas:

- **T1** - Revisão da Literatura;
- **T2** - Experiências iniciais (familiarização com ferramentas e dados a serem utilizados);
- **T3** - Classificação automática de atos de diálogo;
- **T4** - Classificação automática de *intents*;
- **T5** - Descoberta e representação de fluxos de diálogo numa forma gráfica de forma que seja interpretável pelos humanos;

[1] Forma de comunicação com a equipa através de mensagens instantâneas. https://slack.com [março 2023]
[2] Software para realização de videochamadas. https://zoom.us [março 2023]
[3] Serviço de comunicação por vídeo que foi desenvolvido pela Google. https://meet.google.com [março 2023]
[4] Serviço na nuvem que serve para armazenar e sincronizar arquivos. https://drive.google.com [março 2023]
[5] Serviço na nuvem que permite combinar código executável e texto no mesmo documento. https://colab.research.google.com [março 2023]
[6] Linguagem de programação de alto nível. https://www.python.org [março 2023]

- **T6** - Exploração de formas de auxílio a agentes humanos ou inteligentes com recurso aos fluxos descobertos;
- **T7** - Avaliação e análise dos resultados e de soluções;
- **T8** - Escrita de artigo científico;
- **T9** - Escrita do relatório do Projeto.

Na Figura 1.2 está apresentado um diagrama de Gantt[7] que representa um cronograma com a organização das tarefas anteriormente descritas e que foram desenvolvidas ao longo do período do projeto.

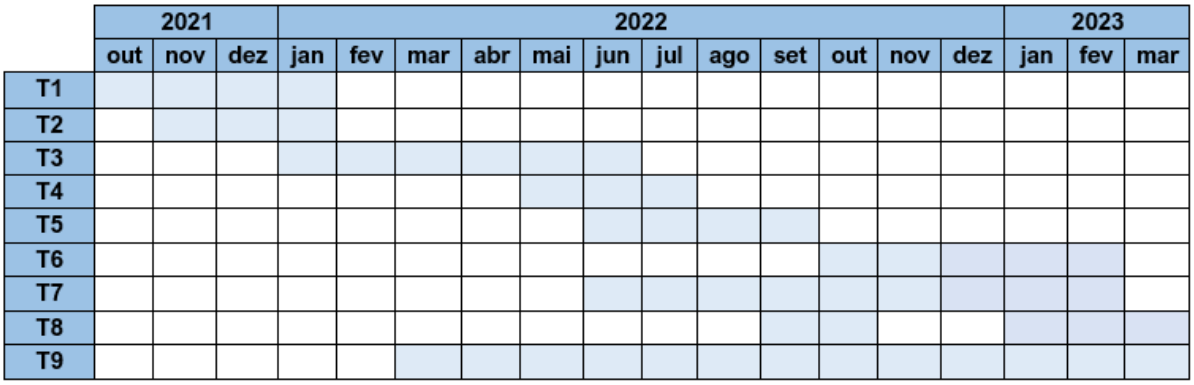


Figura 1.2 - Diagrama de Gantt com as tarefas realizadas representadas

1.5 Contribuições

A elaboração deste projeto permitiu uma série de contribuições diretas, tais como:

1. Abordagem para descobrir e visualizar fluxos a partir de um histórico de diálogos, sem requerer a respetiva anotação, e por isso virtualmente aplicável a qualquer domínio;
2. Estudo acerca da interpretação de *clusters* de falas identificados, com base em diferentes representações e comparação com anotações de atos de diálogo e *intents*;
3. Visualização de fluxos que permitem dar uma visão alto nível das conversas típicas entre humanos, ou humanos e máquinas;
4. Sistemas de Recomendação (*chatbot*) que através do uso de fluxos é capaz de auxiliar centros de atendimento;
5. Elaboração e colaboração em artigos científicos:

[7] Gráfico usado para planear, organizar e ilustrar o avanço de diferentes tarefas num projeto

a. Para a conferência RECPAD 2022 (disponível no Anexo B - Artigo publicado - RECPAD 22). Neste artigo estudou-se se o *clustering* conseguia aproximar-se de *intents* e/ou de atos de diálogo com o uso de algumas técnicas usadas neste trabalho.
b. Coautoria de artigo para a conferência LREC 2022. Foi desenvolvido pela falta de dados reais, identificaram-se vários corpora de diálogo disponíveis de forma gratuita. Fez-se uma avaliação de vários aspetos como o tamanho do conjunto de diálogo, idiomas, anotações e domínios (Gonçalo Oliveira et al., 2022) (disponível no Anexo C – Coautoria artigo publicado - LREC 22).
c. Coautoria num artigo para o RECPAD 2022. Neste trabalho utilizaram-se diversas técnicas de PLN para classificação de atos de diálogo (disponível no Anexo D - Coautoria Artigo publicado - RECPAD 22).
d. Submissão de um artigo científico para a revista Applied Intelligence (APIN) da editora Springer, com o resumo do trabalho presente neste relatório até à experiência no *chatbot*.

1.6 Estrutura do Documento

Este documento é constituído por seis capítulos, estando os mesmos organizados na seguinte forma:

No capítulo 2 são apresentados os diferentes conceitos que são abordados neste relatório de uma forma mais fundamentada, com recurso a trabalhos desenvolvidos nas áreas dos mesmos. Este capítulo serve como base para uma melhor compreensão do que será tratado ao longo deste relatório.

No capítulo 3 é explicada a abordagem seguida ao longo do trabalho e que serviu como base para a exploração e desenvolvimento de técnicas para o projeto. São descritas as diversas especificações necessárias e utilizadas, nomeadamente as técnicas de *Machine Learning* e os conjuntos de dados.

Relativamente ao capítulo 4, são descritos os testes realizados e é feita a avaliação e discussão dos resultados.

No capítulo 5 são apresentados técnicas e testes para tentar representar os fluxos de diálogo partindo dos *clusters* que foram produzidos e a sua aplicabilidade a um *chatbot*.

Por último, no capítulo 6, é feito o balanço do trabalho que foi realizado, o que se atingiu, que desafios trouxe e quais os passos futuros que surgirão das conclusões retiradas.

No fim do documento há uma secção de Anexos com materiais complementares ao Projeto.

2 ESTADO DA ARTE

Neste capítulo é apresentada a pesquisa efetuada sobre os tópicos que estão relacionados com o projeto. Primeiramente, é feita uma contextualização do que são atos de diálogo e de *intents* no panorama do PLN. São dados exemplos de trabalhos realizados de forma supervisionada e não supervisionada na tentativa de descobrir atos de diálogo e *intents*, nomeadamente das técnicas utilizadas e dos resultados obtidos até então por outros autores.

De seguida, identificam-se formas de etiquetar *clusters*. Tratando-se de uma abordagem não supervisionada, serão necessárias técnicas para dar um nome aos *clusters* que são produzidos. Por fim, é feita uma análise aos fluxos de *clusters*, a sua contribuição e como será possível conseguir obtê-los, de forma a auxiliar os agentes humanos ou, criando agentes inteligentes, tornando o atendimento aos clientes autêntico e eficiente.

2.1 Diálogos

Os diálogos são conjuntos de falas e podem ser classificadas como: atos de diálogo, que representam a ação realizada pelo interlocutor (Eisenstein, 2018) (Austin, 1975); ou *intents*, que se traduzem na intenção final do utilizador. Os atos de diálogo são mais genéricos que os *intents*.

A classificação de falas é uma importante área de Inteligência Artificial que tem sido foco de vários trabalhos de investigação. Tal deve-se ao facto de os humanos terem ambição de criar máquinas, agentes inteligentes, capazes de se aproximar o mais possível dos humanos, seja em forma de pensamento ou de expressão linguística.

As falas podem ter diferentes tipos de representação, como *tokens* (unidades individuais de texto), segmentos (grupos de palavras relacionadas) e/ou contexto (informação ao redor das palavras). As falas necessitam de técnicas de *Word Embedding*, ou seja, transformar o texto em vetores numéricos codificados. No caso de (Duran et al., 2021) utilizaram o word2vec (Mikolov et al., 2013) e o GloVe (Pennington et al., 2014), enquanto que (Park et al., 2022) utilizou os modelos de *Sentence Transformer*, o *MiniLM-L6* e o *MiniLM-L12*.

As anotações das falas quanto ao tipo de diálogo tiveram de sofrer uma padronização de forma a uniformizar e, para isso, foi definida a norma ISO 24617-2 (Bunt et al., 2016). De modo simplificado, a ISO mencionada alega que a representação de atos de diálogo deve ser realizada em segmentos individuais/funcionais, em oposição a apenas uma fala, isto pelo facto de uma frase poder conter mais do que um ato de diálogo.

Os atos de diálogo podem e devem ajudar a analisar a linguagem humana para que estes agentes consigam interpretar uma conversa, seja de forma escrita ou falada[8]. Numa conversa, a linguagem que é utilizada na interação entre diferentes oradores, pode provocar mudanças no seu estado cognitivo. Esse efeito de uma fala num determinado contexto, tem o nome de ato de diálogo. Exemplificando, o orador A tem o objetivo de que o orador B faça algo, pode usar diferentes e variadas expressões para alcançar esse objetivo, e todas essas diferentes declarações partilham a mesma função, o ato de diálogo (Popescu-Belis, 2005).

Existem atos de diálogo bastante básicos, tal como uma resposta de sim ou não. No entanto, também existem atos mais complexos, como uma saudação em que uma fala como *"bom dia"* ou *"olá"* seria categorizada com a etiqueta de saudação. Na Tabela 2.1 é possível ver alguns atos de diálogo com a respetiva fala.

Tabela 2.1 - Exemplos de atos de diálogo

Fala	Ato de diálogo
Eu penso que é uma boa opção.	Opinião
Olá, bom dia.	Saudação
A que horas é que parte o comboio?	Interrogativo
Muito obrigado pela ajuda prestada.	Agradecimento
Acho que devias ter mais cuidado.	Sugestão

Por sua vez o *intent* é mais facilmente interpretável que o ato de diálogo, pois categoriza a intenção final para uma fala do utilizador (Truong et al., 2004), e é normalmente mais específico, dependendo sempre do domínio inserido. Na Tabela 2.2 estão apresentadas algumas falas com os respetivos *intents* associados.

Tabela 2.2 - Exemplos de *intents*

Fala	*Intent*
Gostaria de pedir uma sopa e massa bolonhesa como prato principal.	Pedir Comida
Preciso de uma reserva no restaurante OAK para as 20 horas da próxima quinta-feira.	Reservar Restaurante
Olá, procura um hotel no centro de Coimbra para a passagem de ano que tenha wi-fi gratuito.	Procurar Hotel
Boa tarde, sabe dizer-me se existe algum jardim perto do centro da cidade?	Pesquisar Atrações

[8] What Is a Dialog Act? https://www.languagehumanities.org/what-is-a-dialog-act.htm [março 2023]

2.1.1 Atos de Diálogo

Como já mencionado, os atos de diálogo devem ser etiquetados de acordo com a norma ISO 24617-2, que sugere etiquetar os diversos segmentos de uma fala. Várias abordagens foram desenvolvidas para a classificação de atos de diálogo (DAC, acrónimo do termo inglês *Dialog Act Classification*). A maioria adota uma abordagem supervisionada, com modelos treinados em corpora de diálogos onde os atos de diálogo foram anotados manualmente e previamente.

Um destes corpora é o Switchboard-1 Telephone Speech Corpus (Godfrey et al., 1992), um corpus de conversas de texto e de áudio constituído por diversos oradores. O facto de ter sido anotado manualmente é uma vantagem relativamente a outros corpora, pois pressupõe-se uma redução significativa de erros de etiquetagem.

(Stolcke et al., 2000) utilizaram o Switchboard (Godfrey et al., 1992) com o objetivo de classificar automaticamente os 42 atos de diálogo das 1.155 conversas desse corpus. Criaram um sistema através de um modelo hierárquico de três níveis diferentes de classificação utilizando Modelos Ocultos de Markov, conjunto de características lexicais e classificações ao nível da palavra.

Um estudo realizado por (Ribeiro, Ribeiro, & De Matos, 2019) mostrou que é possível obter bons resultados ao classificar atos de diálogo usando uma abordagem de tokenização do texto, usando para isso essas unidades de texto mais pequena, os *tokens*. A abordagem usada pelos autores resume o histórico do diálogo passando a sequência de etiquetas de atos de diálogo por meio de uma GRU (acrónimo do termo em inglês *Gated Recurrent Unit*, uma unidade recorrente fechada, um exemplo de rede neuronal). Essa técnica permite que a rede neuronal "aprenda" as relações entre as diferentes unidades de texto e ajude a identificar os atos de diálogo com mais precisão. Com o modelo, conseguiram obter aplicações a nível da análise de sentimento, classificação de textos ou mesmos sistemas de diálogo.

Outros modelos de *Deep Learning* (Serradilla et al., 2022) podem ser usados neste objetivo, por exemplo diferentes arquiteturas de redes neuronais, tais como as redes neuronais convulsionais (CNN, sigla do termo em inglês *Convolutional Neural Network*) ou recorrentes (RNN, sigla do termo em inglês *Recurrent Neural Network*), ou as memórias de curto prazo longas (LSTM, sigla do termo em inglês *Long Short-Term Memory*) de forma a incorporar informações de contexto para classificação de atos de diálogo (Y. Liu et al., 2017).

Já no que toca à aplicação de aprendizagem não supervisionada, (Ezen-Can & Boyer, 2013) realizaram *clustering* de falas onde os *clusters* obtidos foram avaliados usando um padrão de atos de diálogos já etiquetado manualmente. Foi criada uma matriz das ocorrências de cada ato diálogo em cada *cluster*, e assim avaliar os *clusters* usando as etiquetas predefinidas. Essa matriz de contagens será algo também adotado neste

trabalho, de forma a contabilizar de forma supervisionada a ocorrência dos atos de diálogo ou *intents* em determinados *clusters*.

Através da utilização de *Deep Learning*, nomeadamente CNN e RNN, (Ribeiro, Ribeiro, & Matos, 2019) mostraram que ao prever os atos de diálogo do corpus DIHANA, os diferentes níveis hierárquicos das etiquetas são importantes. Através de árvores de decisão hierárquicas levam em consideração a relação entre diferentes atos de diálogo, e concluindo que a sua previsão de um certo nível é melhorada quando as informações referentes aos níveis superiores estão disponíveis e quando estão incluídas as informações de contexto.

Modelos Ocultos de Markov (Ritter et al., 2010) foram utilizados com informações relativamente a tópicos (*Topic Modeling*) para conversas extraídas do Twitter, ou seja, com diferentes domínios. As palavras dos tópicos foram separadas do restante conteúdo com recurso ao algoritmo *Latent Dirichlet Allocation* (LDA). Esta técnica será adotada e testada neste trabalho. LDA assume que um corpus é um conjunto de tópicos, e estes são definidos por um grupo de palavras que os caracterizam. Este modelo vai receber dois parâmetros, o número de palavras mais significativas a remover e o número de tópicos a que essas palavras estão associadas (X. Li et al., 2015).

Existem trabalhos com abordagens não tão usuais, como é caso de MRF-*based clustering* (MRF, sigla do termo em inglês Markov *Random Field*) (Ezen-Can & Boyer, 2015), em que numa abordagem não supervisionada procurava classificar atos de diálogo de modo a entender melhor as interações de alunos nas aulas, obtendo uma precisão de 36%, algo baixa, mas ainda assim promissora.

2.1.2 Intents

A descoberta de *intents* — por vezes chamado de classificação de *intents* ou intenções — é uma tarefa importante de PLN, que consiste em classificar uma fala quanto ao objetivo principal do utilizador[9].

No contexto de *intents* muitas vezes é confundido o domínio com o tópico, algo que está diretamente relacionado entre si. O domínio é o contexto das falas, os diferentes tópicos que são abrangidos, já o tópico refere-se a um assunto específico dentro de um domínio, enquanto um *intent* refere-se às intenções expressas pelo utilizador. Por exemplo, num domínio de "Serviços bancários", um exemplo de *intent* pode ser "solicitar aumento de crédito". Neste caso, o tópico seria o "aumento de crédito". Dentro de um domínio podem existir diferentes *intents*, bem como diferentes tópicos. Estes representam o assunto específico do utilizador quando expressa uma *intent*.

[9]https://medium.com/mysuperai/what-is-intent-recognition-and-how-can-i-use-it-9ceb35055c4f [março 2023]

Para descobrir *intents*, existem várias abordagens já utilizadas, incluindo o uso de modelos de *Sentence Embedding*, como proposto por (Casanueva et al., 2020). *Word Embedding* e *Sentence Embedding* são técnicas de transformação de texto em vetores numéricos, mas diferem em que representam palavras e falas, respetivamente. Outra abordagem é o *Deep Aligned Clustering* aplicado por (Zhang et al., 2021) para descobrir novos *intents* com a ajuda de dados já pré anotados utilizando o algoritmo K-Means.

Modelos como as redes LSTM também são muito usados. Usando este modelo (Vedula et al., 2020) tentaram prever se uma fala contém uma *intent* e, em seguida, rotular a *intent* na instrução de entrada. O modelo consiste numa LSTM que captura a semântica contextual e o assunto da conversa de acordo com algumas restrições. Eles adaptam uma abordagem de treino para melhorar a robustez e resistência entre os diferentes domínios.

Muitas outras abordagens foram aplicadas e testadas na descoberta de *intents*, nomeadamente a utilização de regras difusas e identificação de padrões para decidir a resposta do sistema ao utilizador (Griol et al., 2021), *Unified Contrastive Learning* (UniCL) (Mou et al., 2022), entre outras.

Aqui foram apresentadas algumas abordagens supervisionadas e não supervisionadas para a descoberta de atos de diálogo e de *intents*. As abordagens supervisionadas requerem um conjunto de dados etiquetados para treinar um modelo, em contrapartida, as abordagens não supervisionadas não necessitam destas etiquetas. Modelos supervisionados como Redes Neuronais (Ribeiro, Ribeiro, & De Matos, 2019) (Liu et al., 2017), principalmente LSTM (Vedula et al., 2020) (Liu et al., 2017), são muito usados, assim como o K-Means (Zhang et al., 2021)(Park et al., 2022) (P. Liu et al., 2021) para modelos não supervisionados.

Em termos de resultados, as abordagens supervisionadas tendem a obter melhores resultados, uma vez que o modelo é treinado já com exemplos etiquetados, no entanto depende sempre do modelo, da quantidade de dados, e da qualidade de etiquetagem. Ter dados rotulados, sejam com atos de diálogo ou *intents*, é uma vantagem, pois é mais fácil avaliar os resultados obtidos, mas num cenário real, nem todos os dados estão etiquetados ou bem etiquetados (muitas das vezes são humanos a etiquetar manualmente). Esta é uma razão para a escolha de abordagens não supervisionadas, pois assim a metodologia é adaptável a qualquer domínio.

2.2 Fluxos e etiquetagem de clusters

Um dos principais problemas das técnicas de *clustering* são as formas de etiquetar os *clusters*, pois num método não supervisionado apenas sabemos as instâncias que pertencem a cada *cluster*, e é com essas informações que podemos etiquetar os *clusters*. Numa tentativa de resolver esse problema, (Penta & Pal, 2021) focaram-se na tarefa de etiquetagem dos *clusters* através da identificação de tópicos e temas relevantes ao texto desses *clusters*. Através de *clustering* e técnicas de extração de palavras-chave encontram os tópicos relevantes de cada *cluster* recorrendo também ao algoritmo

TextRank, e, desta forma, essas palavras-chaves são assim capazes de explicar o conteúdo dos *clusters*, sendo então uma forma de etiquetagem.

Vários trabalhos aplicam abordagens não supervisionadas onde os *clusters* produzidos podem ser usados como nós dos fluxos de diálogo, representados por grafos (Ritter et al., 2010)(J.-L. Bouraoui & Lemaire, 2017)(Joty et al., 2011). (Ritter et al., 2010) utilizaram um corpus de 1.3 milhões de conversas extraídas do Twitter e, através de abordagens não supervisionadas, conseguiram produzir um Modelo Oculto de Markov, que está representado na Figura 2.1. A escolha do nome para os atos de diálogo foi feita de forma intuitiva, algo que não é viável, porque traz muito trabalho humano, algo que pode ser evitado seguindo outras abordagens. Ainda assim, esta cadeia de Markov foi uma das primeiras abordagens na área, e contempla algumas características interessantes para a criação de um modelo deste tipo, como a utilização de *threshold*, de modo a eliminar transições com baixa probabilidade entre atos de diálogo, e também a utilização de dois estados, "START" e o "END", para dar a conhecer como se começa e termina um diálogo nestas conversas do Twitter.

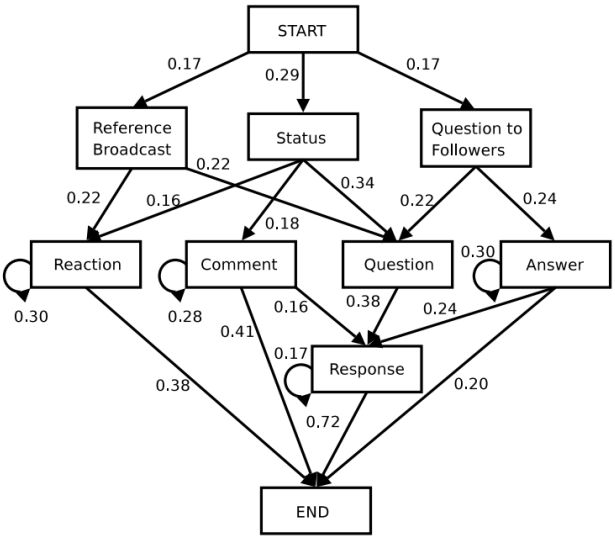


Figura 2.1 - Exemplo de Modelo de Markov com transição entre atos de diálogo (Ritter et al., 2010)

O grafo de transição entre diferentes atos de diálogos presente na Figura 2.1 pode ser um auxílio útil para agentes virtuais e/ou humanos, porque ajuda a prever o desenrolar de um diálogo com um cliente. Ao observar a transição entre os diferentes estados de diálogo, como uma resposta, uma questão, um comentário ou uma reação, pode-se prever o tipo de ato de diálogo quando se recebe determina fala do utilizador, facilitando na rapidez e objetividade de uma resposta. O mesmo se aplica para *intents*, se, por exemplo, o nó é sobre procurar um restaurante, é normal que a transição desse nó seja para outro nó que faça sentido ao domínio em questão.

Estas cadeias de Markov selecionam dinamicamente o próximo movimento numa conversa através da probabilidade de transições entre os diferentes estados.

(Bouraoui & Lemaire, 2017) apresentaram uma proposta usando *clustering* (Ezen-Can & Boyer, 2015) (Zhang et al., 2021) para criar grafos direcionados com as principais transições entre tópicos e diálogos. Essas transições são criadas com recurso a técnicas de *Deep Learning*, Modelos Ocultos de Markov e LDA, com o objetivo de orientar um diálogo, isto é, modelar e orientar as transições entre os diferentes tópicos. Essa orientação entre as transições é feita através da identificação de padrões nas respostas dadas pelo utilizador, identificando o tópico mais provável, e assim selecionar o caminho mais adequado de acordo com a resposta do utilizador.

2.3 Chatbots

Os fluxos de diálogo são importantes porque podem ser aproveitados para sistemas de recomendação ou de guia. Os *chatbots* ou agentes inteligentes são um tipo simples de sistemas de diálogo que podem simular conversas humanas seja via texto ou voz, tendo como principal objetivo fazer querer acreditar que se está realmente a falar com um humano, tornando a conversa o mais autêntica possível.

Graph2Bots (Bouraoui et al., 2019), é uma ferramenta que também visa auxiliar agentes de conversas na extração de grafos de conversas entre humanos. Esta é uma abordagem não supervisionada que utiliza as diferentes fases de diálogo como nós e as suas transições para produzir o grafo.

Gunrock (Chen et al., 2018), um *chatbot* social que ganhou o *Amazon Alexa Award* 2018, mapeia os *intents* dos utilizadores para um determinado tópico, seleciona o módulo mais apropriado para o tópico e encaminha a solicitação do utilizador para ele. Além disso, a fala atual pode ser classificada em tempo real com um ato de diálogo, permitindo a antecipação da próxima fala com diferentes probabilidades, que podem ser usadas para restringir as respostas geradas automaticamente.

Um dos primeiros *chatbots* criados foi o ELIZA (Weizenbaum, 1966), que foi projetado para simular um psicólogo clínico centrado no paciente, ou seja, uma abordagem em que o psicólogo não conhece nada sobre o estado clínico do paciente, portanto a conversa centra-se muito sobre ele, e com perguntas inocentes feitas pelo ELIZA[10], como é possível observar na Figura 2.2.

Muitos *chatbots* tentam passar no teste de Turing[11], e acabam por falhar.

[10] https://web.njit.edu/~ronkowit/eliza.html [março 2023]

[11] É um teste à capacidade de uma máquina exibir comportamento inteligente equivalente a um ser humano https://medium.com/turing-talks/turing-talks-1-o-que-%C3%A9-o-teste-de-turing-ee656ced7b6 [março 2023]

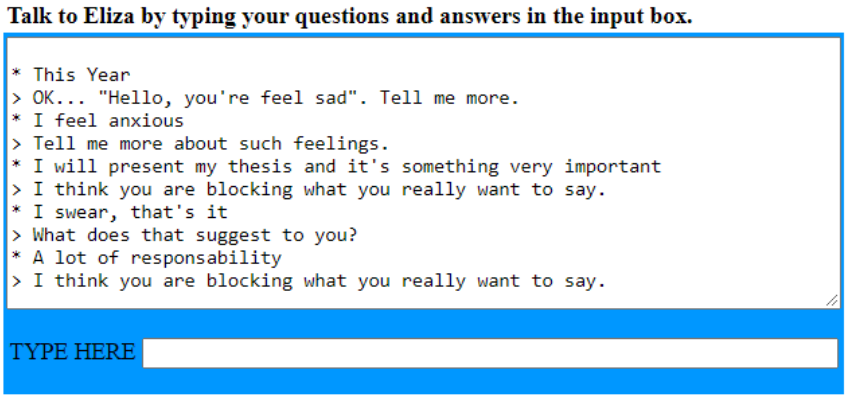


Figura 2.2 - Exemplo de conversa com ELIZA

O primeiro sistema que foi realmente avaliado usando algum teste de Turing foi o PARRY. Este foi desenvolvido para ajudar e falar com pessoas que sofriam de paranoia (Bradeško & Mladenić, 2012).

Estes sistemas inteligentes, podem ter uma aplicabilidade útil em diversas áreas. Por exemplo, na área da saúde, esses sistemas podem ajudar na escolha da medicação de pacientes e até auxiliar no diagnóstico inicial de doenças, segundo (Ayanouz et al., 2020). Este é um exemplo de como um *chatbot* pode ser útil numa área específica.

No entanto, também existem agentes mais generalizados, que podem abordar uma ampla variedade de tópicos (Weitz et al., 2021), tornando o seu domínio o mais abrangente possível. O objetivo utópico é o de tornar o alcance do *chatbot* o mais amplo possível, para que ele possa ser útil em muitos contextos diferentes.

Existem já abordagens destas em português, o *DisBot* (Boné et al., 2020), o primeiro *chatbot* de língua portuguesa a utilizar conhecimento recuperado através dos meios de comunicação para apoiar cidadãos e "socorristas" em cenários de desastres. Ele foi criado com o objetivo de melhorar uma comunidade e a tomada de decisões em cenários de risco.

Diversas técnicas são utilizadas para a criação de um *chatbot*, como uma arquitetura *Encoder-Decoder*[12], que recebe a frase do utilizador como entrada e a converte para um vetor (*Encoder*) e depois produz o output, a frase de resposta (*Decoder*). Esta arquitetura foi utilizada por (Shang et al., 2015) para produzir respostas que estejam enquadradas no mesmo tema de entrada, baseado num treino com um conjunto de dados de perguntas e respostas de um fórum chinês. Apesar de ser uma arquitetura viável, e com alguns bons resultados, principalmente em conversas curtas, elas vão perdendo a coerência à medida que se vão estendendo. Existem muitas outras abordagens que poderiam ser utilizadas na sua criação.

Estes sistemas não têm de ser necessariamente *chatbots*, podem simplesmente fazer recomendações ou obter informações sobre as preferências dos utilizadores. (Yan et al., 2017) identificaram os quatro *intents* mais frequentes num *chatbot* de shopping

[12] https://d2l.ai/chapter_recurrent-modern/encoder-decoder.html [março 2023]

(recomendação, comparação, opinião e pergunta-resposta), conseguindo assim informações úteis para estratégias de Marketing, por exemplo.

Para o desenvolvimento destes agentes podem ser encontradas no mercado diferentes soluções como o *DialogFlow* (Sabharwal & Agrawal, 2020), ou a *Rasa* (Rakesh Kumar Sharma, 2020), entre muitos outros. Qualquer uma destas soluções permite desenvolver um *chatbot* com base num fluxo de diálogo, e a partir daí selecionar as respostas ou ações a tomar. Uma das limitações, no entanto, é a necessidade de criar o fluxo manualmente, o que necessitará sempre de intervenção humana.

2.4 Conclusão

Dada a contextualização de diversos conceitos importantes para a compreensão do trabalho ao longo deste capítulo, na Tabela 2.3 está representada uma uniformização e sumarização dos mesmos.

Tabela 2.3 - Contextualização de conceitos utilizados

Termo	Significado
Corpus	Coleção de texto ou áudio organizado num conjunto de dados
Corpora	Plural de corpus, ou seja, mais que um corpus
Chatbot	Software que tenta simular um ser humano na conversa com outras pessoas.
Fala	Entrada fornecida pelos utilizadores, e considerada a menor unidade nos conjuntos de dados.
Ato de diálogo	É o contexto de uma fala num diálogo, a função da fala.
Intent	É a intenção do utilizador numa determinada fala.

Um diálogo é composto por diversas falas e que podem ser classificadas relativamente aos *intents* ou a atos de diálogo. Esse será um dos contributos deste trabalho, procurar-se-á saber do que se aproxima mais as falas, com base nas diferentes representações utilizadas. Esta será uma abordagem genérica a vários domínios que, dada uma história de diálogos, irá agrupar as falas semelhantes; criar com isso um grafo baseado nos fluxos entre *clusters*; e permitir a visualização desse grafo. Será depois testado num *chatbot* de modo a tentar criar um sistema de apoio/recomendação.

3 METODOLOGIAS E ESPECIFICAÇÕES

Neste capítulo 3 é descrita toda a metodologia que envolve o trabalho realizado. Começando pela forma como foi estruturado o desenvolvimento e respetivos passos para alcançar a solução final, que contempla as ferramentas utilizadas, o plano de trabalho e um esquema da arquitetura proposta. Posteriormente são descritos os corpora que serão utilizados e as respetivas anotações com atos de diálogo e *intents*, bem como todas as técnicas de pré-processamento e de *Word Embedding* utilizadas na preparação dos dados para os inserir nos modelos não supervisionados. As formas de avaliação e validação dos resultados obtidos, e a sua representação de forma a serem úteis para um agente inteligente também são descritos posteriormente neste capítulo. Resumidamente, este capítulo representa a arquitetura que é a base de todo o desenvolvimento utilizado para o projeto final.

3.1 Arquitetura

Serão aplicadas abordagens não supervisionadas a falas para se aproximar de atos de diálogo e/ou *intents*, que idealmente irão fazer com que os *clusters* representem essas mesmas etiquetas. Esses *clusters* podem ser denominados como nós em fluxos de *clusters*, e este é passo pode servir como auxílio a agentes.

Neste capítulo é descrita toda a arquitetura do trabalho que se encontra esquematizada pelo protótipo da Figura 3.1.

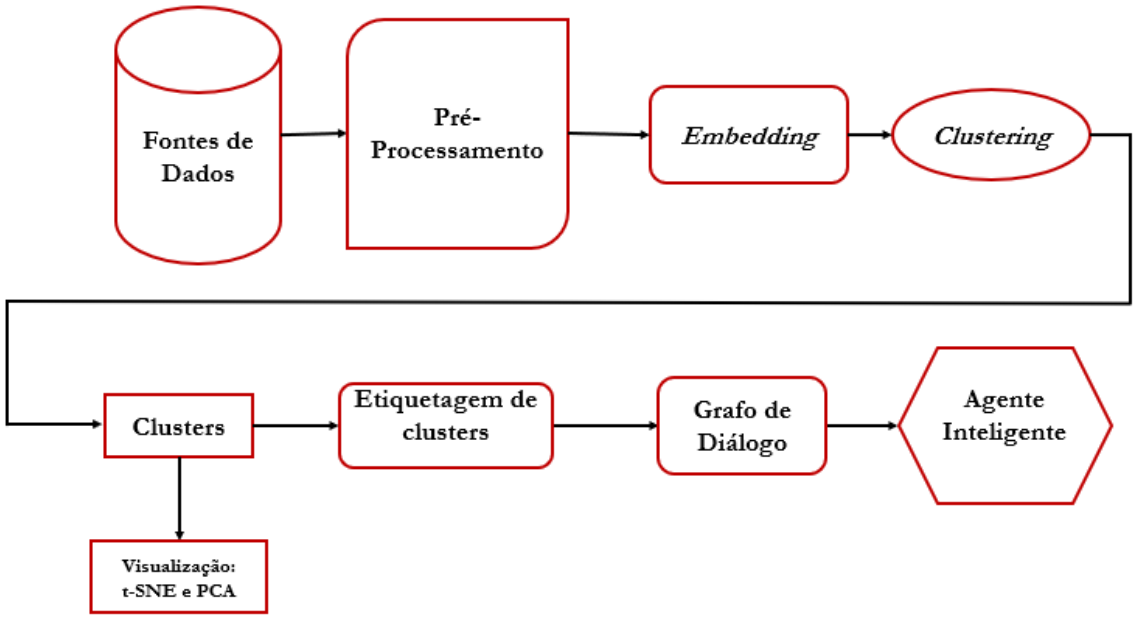


Figura 3.1 - Arquitetura base do trabalho

Com base na Figura 3.1 é possível determinar de forma temporal os passos que foram tomados neste trabalho, e é isso que é explicado neste capítulo. O primeiro

passo é escolher as fontes de dados, os corpora a utilizar, com a mais-valia de terem dados pré-etiquetados para uma melhor avaliação dos resultados. Depois os dados passam por uma série de técnicas de pré-processamento, para tornar os diálogos o mais limpo possível, daí utilizam-se técnicas de Word *Embedding* para transformar o texto em vetores numéricos. Desta forma, é possível através de abordagens não supervisionadas obter *clusters*, que, posteriormente, são processados para que possam ser identificados, ou seja, etiquetados. O passo seguinte é já com as etiquetas associadas aos *clusters*, criar os grafos de diálogo com as diferentes probabilidades e assim auxiliar os agentes inteligentes num diálogo com os clientes.

A elaboração das diversas experiências consiste em três principais etapas: (1) Selecionar e caracterizar os corpora a utilizar, (2) A seleção das técnicas de pré-processamento, na qual engloba as técnicas de *Word Embedding* e (3) a aplicação dos algoritmos não supervisionados com a respetiva avaliação e validação. Este estudo irá englobar a testagem de todas as combinações possíveis das diversas técnicas, algoritmos, com a aplicação nos diferentes corpora, para os problemas de *clustering* de atos de diálogo e de *intents*. Após isso é necessário a aplicação de diferentes técnicas para a etiquetagem dos *clusters* produzidos, pois trata-se de uma abordagem não supervisionada e, portanto, não têm nenhuma etiqueta associada.

3.2 Caracterização dos corpora

De acordo com a Figura 3.1 a primeira etapa consiste em definir que corpora serão utilizados. São todos corpora públicos e com os dados já pré-etiquetados, em que o CamRest676 (Wen et al., 2017), DailyDialog (Y. Li et al., 2017) e Mastodon (Cerisara et al., 2018) são exclusivamente etiquetados com atos de diálogo, enquanto que o SNIPS (Coucke et al., 2018) é um corpus com os *intents* identificados, e os corpora SGD (Rastogi et al., 2020) e MultiWOZ (Budzianowski et al., 2018) têm etiquetas de ambos, atos de diálogo e de *intents*. Para a escolha dos corpora, procurou-se que os dados já estivessem pré-etiquetados. Esta pesquisa contribuiu para a escrita de um artigo científico (Gonçalo Oliveira et al., 2022), que está inserido no projeto em questão, mencionado no subcapítulo 1.2, que resumidamente descreve diversos corpora, aponta os seus pontos fortes e fracos, sendo, portanto, uma pesquisa útil que pode levar investigadores a escolherem sobre determinados conjuntos de dados.

3.2.1 CamRest676

Este corpus *multi-label* é composto por 676 conversas na área da restauração entre um utilizador e uma máquina, com um total de 6069 falas. Os principais atos são o *inform, request, none* e *nooffer*, que se referem respetivamente a atos informativos, solicitações de algo, nenhum ato e um ato de diálogo em que a fala não tem nada para oferecer ao que foi pedido. Este corpus apresenta também o tópico sobre determinada fala, podendo abordar-se assuntos como o preço, área, endereço, entre outros. Assim, pode-se adaptar a um problema de quatro classes, as referidas acima,

ou, especificando com informações de domínio, tornando-se num problema de dezasseis classes, como: *None, Nooffer, Inform-area, Inform-pricerange, Inform-name, Inform-food, Inform-address, Inform-phone, Inform-postcode, Request-area, Request-pricerange, Request-name, Request-food, Request-address, Request-phone* e *Request-postcode*. É possível observar na Figura 3.2 um exemplo de uma conversa completa que inclui alguns dos atos de diálogo.

U: "I want to find a cheap restaurant serving swedish food." - *Inform Food*
S: "I'm sorry, I can't find any cheap Swedish restaurants. Would you like to look for a different style of food, or increase your price range?" - *Request Food*
U: "How about Indian food." - *Inform Food*
S: "There are three cheap Indian restaurants in the center part of town, or Royal spice is a cheap Indian restaurant in the north. Do any of these interest you?" - *Inform Name*
U: "Yes, what is the address for Royal Spice please?" - *Request Address*
S: "Royal spice is located at Victoria Avenue Chesterton. Do you need any other information?" - *Inform Address*
U: "Thank you. Goodbye." - *None*

Figura 3.2 - Exemplo de conversa do CamRest676

Na Tabela 3.1 e na Tabela 3.2 está representada a distribuição de cada etiqueta no CamRest676 para 16 e 4 atos de diálogo respetivamente.

Tabela 3.1 - Distribuição de etiquetas do CamRest676 (16 atos de diálogo)

Etiquetas	Representatividade
None	1642
Nooffer	161
Inform_Area	264
Inform_Pricerange	395
Inform_Name	582
Inform_Food	888
Inform_Address	401
Inform_Phone	123
Inform_Postcode	48
Request_Area	88
Request _Pricerange	64
Request _Name	2
Request _Food	168
Request _Address	182
Request _Phone	365
Request _Postcode	115

Tabela 3.2 - Distribuição de etiquetas do CamRest676 (4 atos de diálogo)

Etiquetas	Representatividade
None	1642
Nooffer	161
Inform	2862
Request	1298

3.2.2 DailyDialog

Um corpus muito utilizado em diferentes trabalhos é o DailyDialog, nomeadamente na descoberta de *intents* ou de sentimentos. É *single-label* composto por 13118 diálogos, com um total de 102968 falas, perfazendo uma média de turnos num diálogo por utilizador de 7.9. Por diálogo são utilizadas em média 114.7 palavras, e uma média por fala de 14.6. A linguagem é escrita por humanos e por isso com pouco ruído, estes diálogos entre dois utilizadores refletem a comunicação do dia-a-dia e por isso abrangem diferentes tópicos.

Este corpus é constituído por apenas quatro atos de diálogo: *Comissive*, *Directive*, *Inform* e *Request*, e a representatividade de cada um encontra-se na Tabela 3.3.

Tabela 3.3 - Distribuição de etiquetas do DailyDialog

Etiquetas	Representatividade
Directive	17293
Question	29424
Inform	46528
Comissive	9723

3.2.3 Mastodon

No que a atos de diálogo diz respeito, por último utilizou-se o Mastodon, que é um corpus *single-label* composto por 2205 falas diferentes relativas a uma rede social, por isso os tópicos são de diferentes domínios. Neste corpus decidiu-se fazer algumas alterações ao conjunto inicial, pois apresentava inúmeros atos de diálogo diferentes, por exemplo inicialmente era composto por: *Statements*, *Agreements*, *Disagrement*; que no corpus original são três atos de diálogo, mas que pertencem a um ato de diálogo maior, o *Inform*, e em casos desses foram todos etiquetados com o ato maior, para não haver uma grande dispersão relativamente a atos de diálogo, passando de 27 atos de diálogo originais para 7 atos de diálogo utilizados, como é possível observar na Tabela 3.4 a sua representatividade de cada uma das etiquetas mapeadas bem como as etiquetas originais do corpus.

Tabela 3.4 - Distribuição de etiquetas do Mastodon

Etiquetas	Representatividade	Etiquetas Originais
Inform	1529	Statement; Agreement; Disagreement; Answer.
Question	328	Propositional question; Set question; Choice question
Request	94	Request;
Other	79	Sympathy; Explicit performative;
Suggestion	53	Suggest;
Greetings	45	Initial goodbye; Initial greetings;
Exclamation	39	Exclamation;
Thanking	38	Thanking; Accept thanking;

As alterações que foram mencionadas neste corpus, foram em concordância com o trabalho original dos autores, que consideravam grupos maiores de atos de diálogo, mas acabaram por separá-los num número maior. Na Tabela 3.5 é possível observar uma conversa, com a respetiva transformação que se fez à etiqueta.

Tabela 3.5 - Exemplo de conversa no Mastodon

ID	Fala	Etiqueta Original	Etiqueta Transformada
0	Does anyone ever feel anxious and empty at the same time?	Propositional question	Question
1	all the time	Answer	Inform
2	i feel like I'm losing my mind a little bit	Statement	Inform
3	relatable. im always anxious and if im not feeling empty or depressed im angry. also usually dissociating.	Agreement	Inform
4	i needa go smoke	Statement	Inform
5	have a good time then! enjoy your smoke.	Initial greetings	Greetings
6	i sure will	Agreement	Inform
7	thank you so much	Thanking	Thanking
8	I hope you start feeling a lil better today	Initial greetings	Greetings

3.2.4 SNIPS

O SNIPS Voice é um corpus *single-label* composto por sete *intents* diferentes. De acordo com a Tabela 3.6, é composto por 13784 diferentes falas, e é um conjunto

de dados bem balanceado nos diferentes *intents*, também é possível observar um exemplo para cada uma das classes representadas.

Tabela 3.6 - Distribuição e exemplo de etiquetas do SNIPS

Etiquetas	Representatividade	Exemplo Frase
Book Restaurant	1881	I want to book a highly rated restaurant for me and my boyfriend tomorrow night.
Get Weather	1896	Is it windy in Boston, MA right now?
Search Screening Event	1852	Check the showtimes for Wonder Woman in Paris
Rate Book	1876	Give 6 stars to Of Mice and Men
Play Music	1914	Play the last track from Beyoncé off Spotify
Add To Playlist	1818	Add Diamonds to my roadtrip playlist
Search Creative Work	1847	Find me the I, Robot television show

3.2.5 MultiWOZ

O Multi-Domain Wizard-of-Oz ou MultiWOZ é um corpus de conversas escritas entre humanos, *multi-label* e que abrange diferentes domínios e tópicos. A versão usada é a 2.2, e pode ser usada para atos de diálogo e *intents*, e é composta por cerca de três mil diálogos. A distribuição das *labels* para atos de diálogo e *intents* pode ser encontrada nas tabelas Tabela 3.8 e Tabela 3.7.

Tabela 3.7 - Distribuição de etiquetas do MultiWOZ para atos de diálogo

Etiquetas	Representatividade
Train Inform	2348
Hotel Inform	1892
Restaurant Inform	1803
Attraction Inform	1516
General Bye	1181
General Thank	936
Train Request	623
Booking Inform	550
Taxi Inform	528
Booking Book	527
Attraction Request	455
Hotel Request	279
Train Offer Booked	272
Restaurant Request	247
General Reqmore	217
None	213
Taxi Request	197
Booking Request	188
Booking No Book	126
Restaurant No Offer	98
Attraction Recommend	96
General Welcome	81
General Greet	72
Train Offer Book	58
Hotel Recommend	56
Hotel No Offer	53
Attraction No Offer	52
Restaurant Recommend	38
Hotel Select	10
Restaurant Select	10
Attraction Select	9
Train No Offer	8
Police Inform	3
Train Select	2

Tabela 3.8 - Distribuição de etiquetas do MultiWOZ para *intents*

Etiquetas	Representatividade
Find Train	1562
Find Hotel	1235
Find Attraction	1232
Find Restaurant	1183
None	1121
Book Restaurant	580
Find Taxi	495
Book Hotel	483
Book Train	431
Find Hospital	12
Find Police	5

O corpus etiquetado com atos de diálogo é composto por 34 *labels*, e para *intents* é composto por 11 *labels*. Como referido é um corpus *multi-label*, e na Figura 3.3 é possível observar uma conversa entre dois utilizadores, onde existem falas com mais que uma etiqueta associada.

```
User1: "I'm looking for a places to go and see during
my upcoming trip to Cambridge." - Find Hotel, Find
Attraction.
User2: "I would like to go to the south area please."
- Find Attraction.
User1: "That's a lot of choices, how about nightclubs?
Can you get me a couple of phone numbers for good
ones? I'm hoping to catch DJ Squalour from Ibiza!" -
Find Attraction.
User2 "Yes, I would like an expensive hotel with free
parking." - Find Hotel.
User1: "Is it four star?" - Find Hotel.
User2: "Okay well now I need to get a taxi to leave
the hotel by 17:45 to the attraction. I will need a
contact number and car type." - Find Taxi.
```

Figura 3.3 - Exemplo de conversa no MultiWOZ

3.2.6 Schema Guided Dialogue

O Schema Guided Dialogue (SGD) são conversas entre humanos e um assistente virtual em que abrange mais de 20 domínios, entre os quais, viagens, tempo, eventos, etc. É um corpus *single-label* e que pode ser usado para classificação de atos de diálogo e de *intents*. A distribuição das *labels* para *intents* e atos de diálogo pode ser encontrada nas tabelas Tabela 3.9 e Tabela 3.10.

Tabela 3.9 - Distribuição de etiquetas do SGD para *intents*

Etiquetas	Representatividade
None	2911
FindProvider	2365
BookAppointment	2354
FindEvents	1863
FindAttractions	1806
ReserveCar	1760
ReserveRestaurant	1754
FindMovies	1744
GetCarsAvailable	1723
FindBus	1604
SearchHotel	1487
BuyEventTickets	1484
FindTrains	1420
FindRestaurants	1366
GetWeather	1309
BuyBusTicket	1263
SearchHouse	1131
PlayMedia	1116
ShareLocation	1077
SearchRoundtripFlights	1057
GetRide	1001
ReserveHotel	931
ScheduleVisit	920
PlayMovie	824
LookupMusic	812
GetTrainTickets	702
FindHomeByArea	663
BookHouse	626
SearchOnewayFlight	592
MakePayment	564
AddAlarm	473
BuyMovieTickets	473
GetAlarms	455
RequestPayment	374
GetTimesForMovie	293

Tabela 3.10 - Distribuição de etiquetas do SGD para atos de diálogo

Etiquetas	Representatividade
Inform	21390
Request	14950
Offer	9426
Confirm	5697
Select	4721
Inform _Intent	4682
Goodbye	4201
Affirm	3397
Thank_You	3111
Notify _Success	2875
Req_More	2684
Offer_ Intent	2260
Negate	1891
Negate _ Intent	1086
Request _Alts	1069
Affirm_Intent	879
Notify_Failure	275

O corpus etiquetado com atos de diálogo é composto por 17 *labels*, e para *intents* é composto por 35 *labels*. Na Tabela 3.11 é possível observar seis falas, 3 exemplos para atos de diálogo e outros três para *intents* do corpus SGD.

Tabela 3.11 - Exemplos de falas para o SGD

Label	Fala
Inform	Hi, could you get me a restaurant booking on the 8th please?
Request	Any preference on the restaurant, location and time?
Offer	I was unable to book your table with those details...
None (*intent*)	No, not right now, thanks for your assistance.
Find Provider	Hi, could you get me a therapist please?
Book Appointment	On which day should i book?

Como a abordagem proposta é não supervisionada, usamos apenas os dados de teste, no caso do MultiWOZ e do SGD.

3.3 Pré-Processamento

Esta fase é resumidamente a preparação dos dados, pois vêm no seu formato original e necessitam da aplicação de algumas técnicas antes de poderem ser utilizados, e foram aplicadas a todos os corpora utilizados.

3.3.1 Técnicas de PLN

Para a preparação das falas extraídas dos corpora são utilizadas técnicas normalmente usadas na área de PLN, tais como, a lematização, que basicamente transforma as palavras do texto na sua forma canónica, a tokenização, que separa as frases nos seus elementos, remoção de pontuação. Também se aplicou a remoção de palavras, como as *stopwords* ou palavras que indiciem um tópico. As *stopwords* são removidas de um dicionário já criado que é composto por palavras consideradas sem utilidade e aplicabilidade num texto, como artigos definidos. A remoção das palavras que indicie um tópico, resume-se a uma forma de tornar o texto o mais generalizado possível, isto numa tentativa de que um texto não se concentre em tópicos, mas sim nos atos de diálogo, que é o que se pretende agrupar como um dos objetivos do projeto.

Aplicou-se também o reconhecimento de Entidades e Nomes (NER, acrónimo do termo inglês *Named Entity Recognition*) que é um modelo pertencente ao spaCy[13] , que serve para reconhecimento de Entidades, e o objetivo principal é de o identificar e classificar as principais informações num texto. Mais concretamente, nomes de pessoas, locais, organizações, datas, entre outras coisas, é visto como menções a entidades, então são delimitadas no texto.

Foi também utilizada uma técnica de pré-processamento que consiste na utilização de *placeholders* para substituir texto que possa ser lixo ou mesmo que possa confundir os modelos, texto esse como *tags* de *usernames* ou URLs. Um exemplo da aplicação prática de cada técnica encontra-se representada na Tabela 3.12.

[13] É uma biblioteca de software de código Python com diversas aplicações. https://spacy.io/ [março 2023]

Tabela 3.12 - Aplicação das Técnicas de Pré-Processamento

Técnica	Frase Original	Frase Processada
Tokenização	Let's go on a trip to Japan.	'Let's, 'go, 'on, 'a', 'trip', 'to', 'Japan'
Remoção de pontuação	Hello! How can I help you?	Hello How can I help you
Lematização	I ate an apple.	I eat a apple.
Remoção de *stopwords*	I am a good boy.	good boy.
Topic Modeling	This restaurant serves really good steak.	This serves really good. (remoção de palavras relativas ao tópico de restauração)
Substituição URL	You can access the desired link here: https://spacy.io/	You can access the desired link here: xURLx
Substituição *usernames*	@DLMartins we will put you in conversation with one of our agents.	xUSERx we will put you in conversation with one of our agents.
NER	Daniel Martins wanted to do a Google search on New York in November 2022.	PERSON wanted to do a ORG search on LOCATION in DATA

Diferentes técnicas acima, foram inseridas numa grelha de parâmetros, do estilo *Grid Search*[14], que será explicada e apresentada nos diferentes testes nos capítulos seguintes, isto com o objetivo de ver se as aplicabilidades das técnicas teriam efeitos positivos ou não nos resultados obtidos.

Uma dessas técnicas foi a de utilização de *Topic Modeling*, uma sugestão feita pelos autores (Ritter et al., 2010), para isso foi utilizada a biblioteca NLTK[15] e com recurso a LDA[16], uma abordagem para a modelação de tópicos que resumidamente considera cada documento como uma coleção de tópicos com determinada representatividade. É dado o número de tópicos principais, e o algoritmo reorganiza os dados e a distribuição dos tópicos, e encontra as palavras-chaves que determinar os tópicos e foram essas as palavras que foram eliminadas.

Apesar da aplicação das técnicas anteriormente descritas, o corpus ainda está em formato textual, apenas com algumas modificações em virtude das técnicas. Portanto, é indispensável que se transforme o texto em vetores numéricos, isto de

[14] Em Machine Learning, é a pesquisa exaustiva para obter os melhores parâmetros
[15] Conjunto de bibliotecas da área de PLN. https://www.nltk.org/ [março 2023]
[16] Modelo estatístico generativo

modo que se possa aplicar algoritmos de aprendizagem não supervisionada. Para isso utiliza-se *Word Embeddings*, que é uma abordagem para representar palavras e documentos em formato de vetores numéricos. Isto servirá como *input* para os modelos que serão posteriormente usados.

3.3.2 Word Embedding

Serão utilizadas três diferentes técnicas de *Word Embeddings*, nomeadamente:

O word2vec (Mikolov et al., 2013) é um método para processar texto e "vetorizar" palavras. É um modelo composto por uma rede neuronal com duas camadas com uma camada de entrada, uma camada oculta e uma camada de saída, que durante o treino vai ajustar os pesos para reduzir a função de *loss*. Recebe como entrada um texto pertencente a um corpus, e produz como saída um conjunto de vetores, que representam as diferentes palavras do corpus. A utilidade deste método é o facto de agrupar os vetores de palavras semelhantes próximos entre si no espaço vetorial, isto é, deteta as semelhanças entre as palavras de forma matemática.

TF-IDF (Jones, 1972)(Luhn, 2010) é uma técnica que combina termo e frequência, que resumidamente dá-nos a frequência de cada termo numa determinada frase, e assim cada palavra terá determinado peso, conforme a sua representatividade (frequência), e esses valores são guardados num vetor[17].

Sentence Transformers (Reimers & Gurevych, 2019) são uma evolução do *Bidirectional Encoder Representations from Transformers* (BERT). O BERT foi projetado para entender e codificar palavras e frases, enquanto o *Sentence Transformer* tem como objetivo codificar falas ou diálogos inteiros. Esta *framework* é composta por diferentes modelos pré-treinados e destinados a diferentes tarefas.

Neste trabalho foram realizados testes com quatro versões de modelos diferentes[18], nomeadamente *all-mpnet-base-v2*, *all-distilroberta-v1*, *all-MiniLM-L12-v2* e *all-MiniLM-L6-v2,* estes modelos foram escolhidos através de uma extensiva avaliação[19] que utilizou diferentes modelos para diferentes objetivos e foram escolhidos os quatro modelos que obtiveram melhores resultados para a performance de *Sentence Embeddings*, que seria esse o objetivo. Os modelos *all-mpnet-base-v2* e *all-distilroberta-v1*, fazem o mapeamento das diferentes frases para um espaço vetorial de 768 dimensões, enquanto o *all-MiniLM-L12-v2* e o *all-MiniLM-L6-v2* o fazem para um espaço de 384 dimensões. Algo a avaliar durante este trabalho, será o impacto dos diferentes modelos usados.

[17] https://towardsdatascience.com/how-tf-idf-works-3dbf35e568f0 [março 2023]
[18] Os modelos são obtidos em: https://huggingface.co/ [março 2023]
[19] Essa avaliação foi realizada por:
https://www.sbert.net/docs/pretrained_models.html#sentence-embedding-models/ [março 2023]

3.4 Aprendizagem Não Supervisionada

Após a aplicação de *Word Embeddings*, os algoritmos de ML conseguem processar os vetores numéricos que lhes são dados como *input*, ao invés de texto. Para que com recurso à aprendizagem não supervisionada, mais precisamente de *clustering*, se tente agrupar os atos de diálogos e/ou *intents* no respetivo *cluster*.
Será utilizado o K-Means (MacQueen, 1967), que dado um grupo de *clusters* inicial, não ideal, irá realocar cada instância para o ponto do centro mais próximo, agrupando assim em *clusters* as instâncias mais semelhantes entre si. Tem um parâmetro K que remete para o número de *clusters* a serem agrupados, este valor será definido de acordo com o número das classes do corpus utilizado.

3.5 Avaliação e Validação dos dados

É importante a etiquetagem dos dados, para que se possa utilizar essas etiquetas e avaliar de forma externa, uma forma de validação dos dados para além da avaliação interna que não requer a etiquetagem. Neste trabalho a avaliação está dividida em dois grupos, uma avaliação interna e externa. Resumidamente a avaliação interna foca-se nos dados que estão presentes nos *clusters*, ou seja, faz uma avaliação relativamente aos atos de diálogo / *intents* de cada *cluster* verificando a semelhança entre cada instância. A avaliação externa avalia quão bem os *clusters* que foram produzidos de acordo com as classes padrão, portanto esta avaliação requer a existência de dados já pré-etiquetados. Para a avaliação interna será utilizado o *Davies-Bouldin Index* (DBI) (Davies & Bouldin, 1979), que basicamente avalia quão bem o *clustering* foi feito, usando quantidades e características inerentes ao conjunto de dados e é a medida de similaridade média de cada *cluster* com o *cluster* mais semelhante. Pode variar entre qualquer valor positivo, e quanto mais próximo do 0 se aproximar, melhor são os indicadores do *clustering*.

Este é definido tem como base a similaridade média entre cada *cluster* C_i para i=1,...,k e o seu *cluster* C_j mais similar. Neste sentido, essa similaridade é definida como uma medida R_{ij}, na qual s_i corresponde à distância média entre cada ponto do *cluster i* e o seu centroide, e o d_{ij} a corresponder à distância entre os centroides dos *clusters* i e o j. A equação da medida R_{ij} pode ser dada pela Equação (3.1)[20]:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{3.1}$$

Portanto a *Davies-Bouldin Index* pode ser definido pela Equação (3.2)[20]:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij} \tag{3.2}$$

Na avaliação interna também é usado o *Silhouette Coefficient* (Rousseeuw, 1987a) que vai comparar a distância média para elementos no mesmo *cluster* com a distância média para elementos noutros *clusters*, o melhor valor é 1 e o pior valor é -1, e em que valores próximos de 0 indicam *clusters* sobrepostos. Este coeficiente resumidamente avalia quão bem os *clusters* foram definidos, e é definido pela Equação (3.3)[21].

$$s = \frac{b - a}{max\ (a, b)} \tag{3.3}$$

Onde o a corresponde à distância média entre uma amostra s e todos os outros pontos pertencentes à mesma classe. Enquanto o b é a distância média entre uma amostra s e todos os outros pontos pertencentes ao *cluster* mais próximo. Depois é feita a média do *Silhouette Coefficient* de todas as amostras, e dada a equação conclui-se que o score será melhor conforme os *clusters* sejam bem separados uns dos outros.

Relativamente à avaliação externa, serão utilizadas duas métricas comumente usadas, que é a *Accuracy* que avalia se as *etiquetas* que foram preditas correspondem às originais, varia entre 0 e 1, e quanto mais próximo de 1 melhor resultado. É utilizada a accuracy_score[21] que devolve a fração das previsões que foram feitas corretamente. Será usado também a *V-Measure* (Rosenberg & Hirschberg, 2007) que corresponde à média harmónica entre a homogeneidade e o grau de completude dos *clusters*, os valores variam entre 0 e 1, em que 1 tende para um *clustering* perfeito (Vysala & Gomes, 2020). A fórmula que calcula a *V-Measure* é dada pela Equação (3.4) [21]:

$$v = \frac{(1 + \beta)\ \text{x homogeneity x completeness}}{(\beta\ \text{x homogeneity}\ +\ \text{completeness}\ )} \tag{3.4}$$

Para os cenários de *multi-label* a avaliação é exata quando a classe predita corresponde exatamente à classe verdadeira.

Em que o valor β é 1 por defeito, e a homogeneidade dada pela Equação (3.5)[21]e é uma medida útil para verificar se o *clustering* cumpriu um requisito importante: conter apenas amostras pertencentes a uma única classe.

[20] Equações e dados retirados de: https://scikit-learn.org [março 2023]
[21] Equações e dados retirados de: https://scikit-learn.org [março 2023]

$$h = 1 - \frac{H(C|K)}{H(C)} \tag{3.5}$$

Em que $H(C|K)$ é a entropia condicional das classes dadas as atribuições do *cluster* e o $H(C)$ é a entropia das classes.

Já o grau de complemento (*completeness*) remete para se todas as instâncias que pertencem a uma classe são elementos do mesmo *cluster*, e é dada pela Equação (3.6)[21]:

$$h = 1 - \frac{H(K|C)}{H(K)} \tag{3.6}$$

Em que $H(K|C)$ é a entropia condicional dos *clusters* dadas as diferentes classes e p $H(C)$ é a entropia dos *clusters*.

Como já referido a avaliação externa pressupõe a existência dos dados já etiquetados com atos de diálogo e/ou *intents*.

3.5.1 Visualização dos Dados

Serão também utilizadas duas técnicas de visualização de dados que consiste na redução de dimensionalidade de forma não supervisionada, são elas o PCA (PCA, acrónimo do termo inglês *Principal Component Analysis*)(Jollife & Cadima, 2016) e o t-SNE (t-SNE acrónimo do termo inglês *t-distributed Stochastic Neighbor Embedding*) (Van Der Maaten & Hinton, 2008). Estas técnicas são usadas tendo em conta a composição e o *clustering* dos dados com dimensões muito elevadas, reduzindo-as para duas dimensões permitindo assim uma visualização eficaz.

O PCA caracteriza-se por identificar as dimensões sobre as quais os dados se encontram mais dispersos, e assim identificar quais os componentes principais do conjunto de dados. Esta técnica pode ser dividida em cinco passos:

1. Padronizar o intervalo de variáveis iniciais contínuas;
2. Calcular a matriz de covariância para identificar as correlações;
3. Calcular os autovetores e autovalores da matriz de covariância para identificar os principais componentes;
4. Criar um vetor de *features* para decidir que principais componentes manter;
5. Reformular os dados ao longo dos eixos.

Esta técnica permite eliminar a redundância entre os dados, bem como representá-los graficamente de forma comprimida. A ideia principal do PCA é simples, é a de

reduzir o número de variáveis do corpus, e preservar ao máximo a informação possível[22].

Enquanto o método t-SNE resume-se a encontrar conexões não lineares entre os dados. E parte de três princípios base para isso:

1. Calcular uma distribuição de probabilidade conjunta que representa as semelhanças entre as instâncias;
2. Criar um conjunto de instância, calculando a distribuição de probabilidade conjunta para eles também;
3. Usa o método do gradiente para alterar o conjunto de instância no espaço de baixa dimensão.

Resumidamente este procedimento probabilístico mapeia as informações multidimensionais para um espaço de menor dimensão e procura descobrir desenhos nos dados[23].

3.6 Etiquetagem de Clusters e Grafo de Diálogo

Após a obtenção dos *clusters* por uma das abordagens não supervisionadas referidas no subcapítulo 3.4, os *clusters* necessitam de ser etiquetados, ou seja, atribuir-lhes uma etiqueta, dar-lhes um nome que não seja simplesmente *Cluster* 0, *Cluster* 1, etc., para que cada *cluster* tenha uma identidade própria.
Para isso foram utilizadas três abordagens diferentes na etiquetagem dos *clusters*.
A primeira abordagem foi a do uso de bigramas (Collins, 1996), ou seja, uma sequência de duas palavras adjacentes numa frase. Então, foram observadas todas as falas presentes em cada *cluster* produzido, e verificado qual o bigrama mais representado em cada *cluster*, e esse bigrama torna-se na etiqueta do *cluster* correspondente. A frase "Eu gosto de batatas fritas" é composta por 4 bigramas, nomeadamente "Eu gosto", "gosto de", "de batatas" e "batatas fritas".
A segunda abordagem é composta pelo uso do verbo principal do sintagma verbal e os seus complementos principais, que é uma estrutura cujo núcleo é um verbo, para etiquetar os *clusters*. Um exemplo pode ser visto nesta frase "A nossa família costuma oferecer gorjeta aos atendentes eficientes", o predicado seria "costuma oferecer gorjeta aos atendentes eficientes", que é constituído pelo verbo principal cujo flema é "oferecer", os complementos "gorjeta" e "atendente", que na plataforma criada, após remoção de *stopwords* e da lematização, a etiqueta fica "oferecer gorjeta atendente", tornando esta numa ação que seria considerada uma intenção dos

[22] https://builtin.com/data-science/step-step-explanation-principal-component-analysis [março 2023]
[23] https://towardsdatascience.com/t-sne-clearly-explained-d84c537f53a; [março 2023]
https://medium.com/swlh/t-sne-explained-math-and-intuition-94599ab164cf [março 2023]

utilizadores. Esta abordagem tem o mesmo princípio da primeira, em que vai ver em todas as falas de cada *cluster* qual o verbo + sintagma verbal mais representado.
Se em alguma das abordagens, existirem *cluster* com a etiqueta repetida, simplesmente acrescenta-se um número que vai incrementando.
A última abordagem, e a menos utilizada neste trabalho, é a de descobrir qual a fala mais próxima do centroide do *cluster*, que é resumidamente o centro de cada *cluster*. Isto faz com que essa fala seja a mais distintiva relativamente às outras e aos outros *clusters*. Resumidamente calcula as distâncias mínimas entre um ponto e o resto dos pontos desse *cluster*[24], e devolve o valor mínimo, ou seja, o do centroide.

A elaboração dos grafos foi feita com recurso a Cadeias de Markov, estas são desenvolvidas tendo em conta os *clusters* que são produzidos e como estão etiquetados. Inicialmente é necessário criar uma matriz de transições, que basicamente converte em probabilidades as transições entre *clusters*, sejam diferentes, ou o mesmo *cluster*.

Depois, com recurso ao NetworkX (Hagberg et al., 2008), que é um package de Python para criar e manipular estruturas como grafos e outro tipo de gráficos, são criados os nós, que correspondem a todos os *clusters*, depois as arestas, de um *cluster* para todos, que corresponde aos valores da matriz de transição.

Assim como em (Ritter et al., 2010), vai ser definido um valor de *threshold*, sempre utilizado como 0.15, para remover as arestas em que a probabilidade é inferior a esse valor, de modo a não tornar a cadeia de Markov demasiado confusa e extensa. O valor 0.15 foi escolhido de forma empírica, após a realização de diversos testes verificou-se que seria o mais adequado de maneira que permanecessem sempre pelo menos uma aresta, as arestas mais importantes e de forma a tornar a cadeia de Markov visualmente interpretável. Deste trabalho será replicada outra abordagem, nomeadamente a utilização de dois *clusters* para o início e fim de diálogo de forma a tornar o fluxo mais interpretável e coerente.

3.7 Arquitetura Final

Após a explicação de todas as metodologias e técnicas utilizadas neste projeto ao longo deste capítulo, é possível agora observar a arquitetura inicial no subcapítulo 3.1 de forma instanciada na Figura 3.4 com todos os passos e escolhas que serão tomados.

[24] É feito com recurso à técnica pairwise_distances_argmin_min

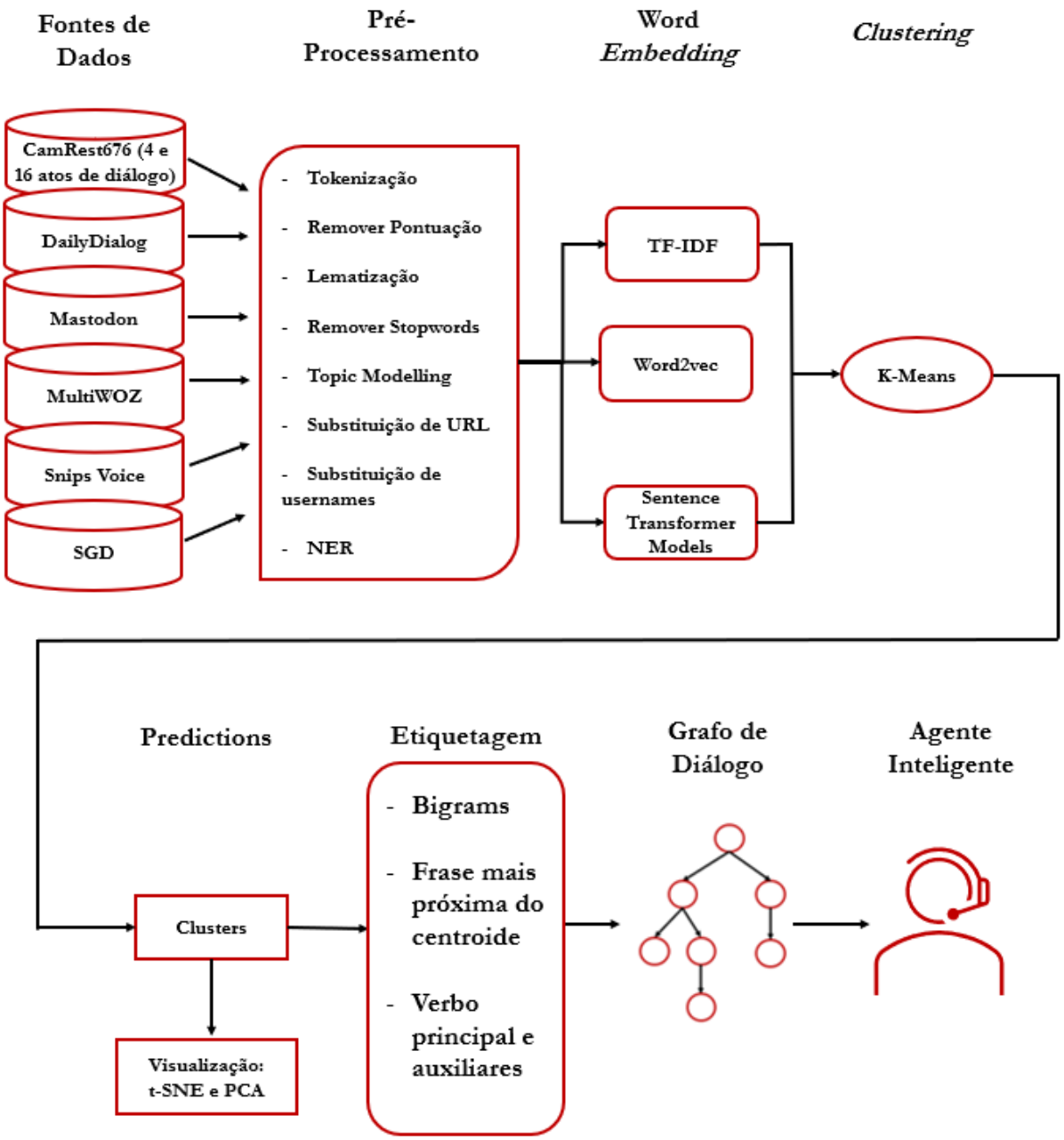


Figura 3.4 - Arquitetura final de trabalho

3.8 Plataforma Final

Esta plataforma desenvolvida cumpre os seguintes requisitos:

1. Aplicação de diferentes técnicas de pré-processamento, e de *Word Embedding* para representação dos dados;
2. *Clustering* e respetiva avaliação (interna e externa) para corpus de diferentes domínios, e problemas *single-label* ou *multi-label*;
    a. Caso seja um corpus não etiquetado (sem *labels*) aplica-se a heurística *Silhouette Method* (Rousseeuw, 1987b) para descobrir o número ideal de *clusters*;
    b. Caso seja um corpus etiquetado, o parâmetro K é escolhido tendo em conta o número exato de atos de diálogo ou *intents* do corpora. De forma a comparar entre si o melhor resultado por avaliação interna/externa;
3. Visualização dos dados obtidos, através de técnicas de redução de dimensionalidade;
4. Produção de fluxo de diálogo com base nos *clusters* obtidos e no método de etiquetagem escolhido;
5. Interação com um *chatbot* tendo em conta o fluxo de diálogo obtido.

4 AVALIAÇÃO E DISCUSSÃO

Neste capítulo serão testados os diferentes métodos propostos no capítulo 3 sobre os corpora descritos anteriormente na tentativa de compreender do que se aproximam mais, se de atos de diálogo se de *intents*.

4.1 Clustering para descobrir Atos de Diálogo

Neste subcapítulo são descritas todas as experiências feitas na tentativa de agrupar os atos de diálogo em *clusters*, para isso foi essencial o estudo dos conceitos apresentados no capítulo 2 e todos os detalhes do desenvolvimento do trabalho referidos no capítulo 3.

4.1.1 Modelos Sentence Transformer

Foram testados os quatro modelos de *Sentence Transformer*, *all-mpnet-base-v2*, *all-distilroberta-v1*, *all-MiniLM-L12-v2* e *all-MiniLM-L6-v2*, de forma a analisar que modelo será utilizado futuramente.

Tabela 4.1 - Variações nos quatro modelos de *Sentence Transformer* aplicados ao CamRest676 de 4 *clusters*

Modelo *Sentence Transformer*	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
all-MiniLM-L6-v2	0.16	2.526	0.087	0.361
all-MiniLM-L12-v2	**0.179**	2.477	**0.137**	0.372
all-distilroberta-v1	0.158	2.494	0.108	0.365
all-mpnet-base-v2	0.157	**2.429**	0.122	0.357

Tabela 4.2 - Variações nos quatro modelos de *Sentence Transformer* aplicados ao CamRest676 de 16 *clusters*

Modelo *Sentence Transformer*	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
all-MiniLM-L6-v2	0.168	2.975	0.044	0.153
all-MiniLM-L12-v2	0.184	3.086	0.056	0.184
all-distilroberta-v1	0.18	2.689	**0.121**	0.214
all-mpnet-base-v2	**0.197**	**2.462**	0.101	**0.263**

Tabela 4.3 - Variações nos quatro modelos de *Sentence Transformer* aplicados ao Mastodon

Modelo *Sentence Transformer*	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
all-MiniLM-L6-v2	0.019	5.545	0.107	0.055
all-MiniLM-L12-v2	0.023	5.432	0.119	0.059
all-distilroberta-v1	0.026	5.513	0.103	0.062
all-mpnet-base-v2	0.021	5.498	0.145	0.056

Tabela 4.4 - Variações nos quatro modelos de *Sentence Transformer* aplicados ao DailyDialog

Modelo *Sentence Transformer*	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
all-MiniLM-L6-v2	0.018	5.557	0.112	0.053
all-MiniLM-L12-v2	0.024	5.439	0.121	0.061
all-distilroberta-v1	0.027	5.508	0.099	0.064
all-mpnet-base-v2	0.02	5.485	0.139	0.055

Tabela 4.5 - Variações nos quatro modelos de *Sentence Transformer* aplicados ao MultiWOZ de atos de diálogo

Modelo *Sentence Transformer*	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
all-MiniLM-L6-v2	0.079	2.882	0.588	0.483
all-MiniLM-L12-v2	0.079	2.83	0.615	0.512
all-distilroberta-v1	0.069	2.978	0.64	0.531
all-mpnet-base-v2	0.076	2.929	0.622	0.51

Tabela 4.6 - Variações nos quatro modelos de *Sentence Transformer* aplicados ao SGD de atos de diálogo

Modelo *Sentence Transformer*	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
all-MiniLM-L6-v2	0.076	3.06	0.4	0.347
all-MiniLM-L12-v2	0.079	3.022	0.401	0.348
all-distilroberta-v1	0.078	3.049	0.395	0.34
all-mpnet-base-v2	0.079	3.057	0.404	0.356

Analisando as tabelas Tabela 4.1, Tabela 4.2, Tabela 4.3, Tabela 4.4, Tabela 4.5 e Tabela 4.6 conclui-se que comparando os quatro modelos, não há nenhum modelo que tenha obtido resultados de destaque, e por isso daqui em diante será utilizado o modelo *all-MiniLM-L6-v2*, pela razão de ser o modelo mais rápido dos quatro, e para servir como base dos modelos de *Sentence Transformer* para comparar com as outras técnicas de *Word Embedding* utilizadas neste trabalho.

Dos resultados obtidos pode-se destacar os elevados valores obtidos de *Accuracy* e *V-Measure* para o MultiWOZ e SGD, comparativamente com os outros corpora. O CamRest676, seja para quatro, ou para dezasseis classes foi o que apresentou os melhores valores de avaliação interna.

4.1.2 TF-IDF

Nesta seção será testada a técnica de *Word Embedding*, o TF-IDF. Inicialmente será avaliado a utilização de diferentes valores de parâmetros, irão ser testadas diferentes combinações para *max-df* e *min-df*. Foram experimentados diferentes valores dos parâmetros *max-df* e *min-df*, respetivamente, o máximo e a proporção mínima onde determinado termo pode ocorrer nas diferentes falas. Por exemplo, *max-df* a 0.8 indica que os termos que apareçam em 80% ou mais das falas não serão incluídos na vectorização, útil para remover palavras demasiado comuns, o que não traria informação importante, e seria algo tendencioso. O parâmetro *min-df* com o valor de 10, indica que os termos que apareçam em menos de 10 falas não serão incluídos na vectorização, algo útil para a remoção de ruído.

Tabela 4.7 - Variações nos parâmetros do TF-IDF aplicados ao corpus CamRest676 de *clusters*

Parâmetros/Métricas	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
max-df:0.75; min-df:05	0.105	3.253	0.087	0.137
max-df:0.80; min-df:05	0.112	3.145	0.103	0.142
max-df:0.85; min-df:05	0.098	3.521	0.094	**0.156**
max-df:0.75; min-df:10	0.126	3.362	0.098	0.151
max-df:0.80; min-df:10	0.124	3.021	**0.113**	0.149
max-df:0.85; min-df:10	**0.127**	**2.961**	0.110	0.154
max-df:0.75; min-df:15	0.119	3.142	**0.113**	0.139
max-df:0.80; min-df:15	0.121	3.057	0.103	0.143
max-df:0.85; min-df:15	0.118	3.239	0.110	0.147

Tabela 4.8 - Variações nos parâmetros do TF-IDF aplicados ao CamRest676 de 16 *clusters*

Parâmetros/Métricas	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
max-df:0.75; min-df:05	0.134	3.415	0.048	0.157
max-df:0.80; min-df:05	0.163	3.614	0.049	0.148
max-df:0.85; min-df:05	0.118	4.257	0.046	0.154
max-df:0.75; min-df:10	0.162	3.433	0.041	0.134
max-df:0.80; min-df:10	0.167	3.122	**0.051**	0.158
max-df:0.85; min-df:10	**0.173**	**2.975**	0.044	0.153
max-df:0.75; min-df:15	0.148	3.515	**0.051**	0.152
max-df:0.80; min-df:15	0.162	3.152	0.049	0.145
max-df:0.85; min-df:15	0.147	3.328	0.05	**0.159**

Tabela 4.9 - Variações nos parâmetros do TF-IDF aplicados ao Mastodon

Parâmetros/Métricas	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
max-df:0.75; min-df:05	0.022	5.523	0.112	0.059
max-df:0.80; min-df:05	0.025	5.489	0.129	0.054
max-df:0.85; min-df:05	0.018	5.527	0.094	0.055
max-df:0.75; min-df:10	0.024	5.52	0.117	0.057
max-df:0.80; min-df:10	**0.026**	**5.414**	0.113	**0.064**
max-df:0.85; min-df:10	0.019	5.534	**0.124**	0.061
max-df:0.75; min-df:15	0.021	5.831	0.105	0.062
max-df:0.80; min-df:15	0.021	5.482	0.109	0.059
max-df:0.85; min-df:15	0.018	5.715	0.103	0.061

Tabela 4.10 - Variações nos parâmetros do TF-IDF aplicados ao DailyDialog

Parâmetros/Métricas	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
max-df:0.75; min-df:05	0.012	5.632	0.082	0.042
max-df:0.80; min-df:05	0.016	5.702	0.091	0.045
max-df:0.85; min-df:05	0.014	5.659	0.097	0.048
max-df:0.75; min-df:10	0.013	5.516	0.101	0.056
max-df:0.80; min-df:10	0.017	5.492	0.098	**0.062**
max-df:0.85; min-df:10	0.019	5.592	**0.106**	0.059
max-df:0.75; min-df:15	**0.023**	5.489	0.102	0.06
max-df:0.80; min-df:15	0.019	**5.463**	0.099	0.057
max-df:0.85; min-df:15	0.017	5.728	0.094	0.058

Tabela 4.11 - Variações nos parâmetros do TF-IDF aplicados ao MultiWOZ de atos de diálogo

Parâmetros/Métricas	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
max-df:0.75; min-df:05	0.052	3.929	0.476	0.366
max-df:0.80; min-df:05	0.054	**3.717**	0.469	0.363
max-df:0.85; min-df:05	0.05	3.967	0.476	0.362
max-df:0.75; min-df:10	0.053	3.788	0.495	0.376
max-df:0.80; min-df:10	0.054	3.863	0.482	0.372
max-df:0.85; min-df:10	0.05	3.786	**0.504**	**0.389**
max-df:0.75; min-df:15	0.056	3.826	0.481	0.374
max-df:0.80; min-df:15	0.055	3.814	0.457	0.35
max-df:0.85; min-df:15	**0.057**	3.77	0.462	0.36

Tabela 4.12 - Variações nos parâmetros do TF-IDF aplicados ao SGD de atos de diálogo

Parâmetros/Métricas	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
max-df:0.75; min-df:05	0.053	4.134	0.409	0.288
max-df:0.80; min-df:05	0.054	4.13	**0.426**	**0.3**
max-df:0.85; min-df:05	0.053	4.134	0.408	0.289
max-df:0.75; min-df:10	0.055	4.183	0.409	0.291
max-df:0.80; min-df:10	0.054	4.176	0.403	0.283
max-df:0.85; min-df:10	0.055	4.064	0.411	0.29
max-df:0.75; min-df:15	0.056	**3.951**	0.405	0.291
max-df:0.80; min-df:15	0.056	4.094	0.412	0.288
max-df:0.85; min-df:15	**0.057**	4.003	0.41	0.292

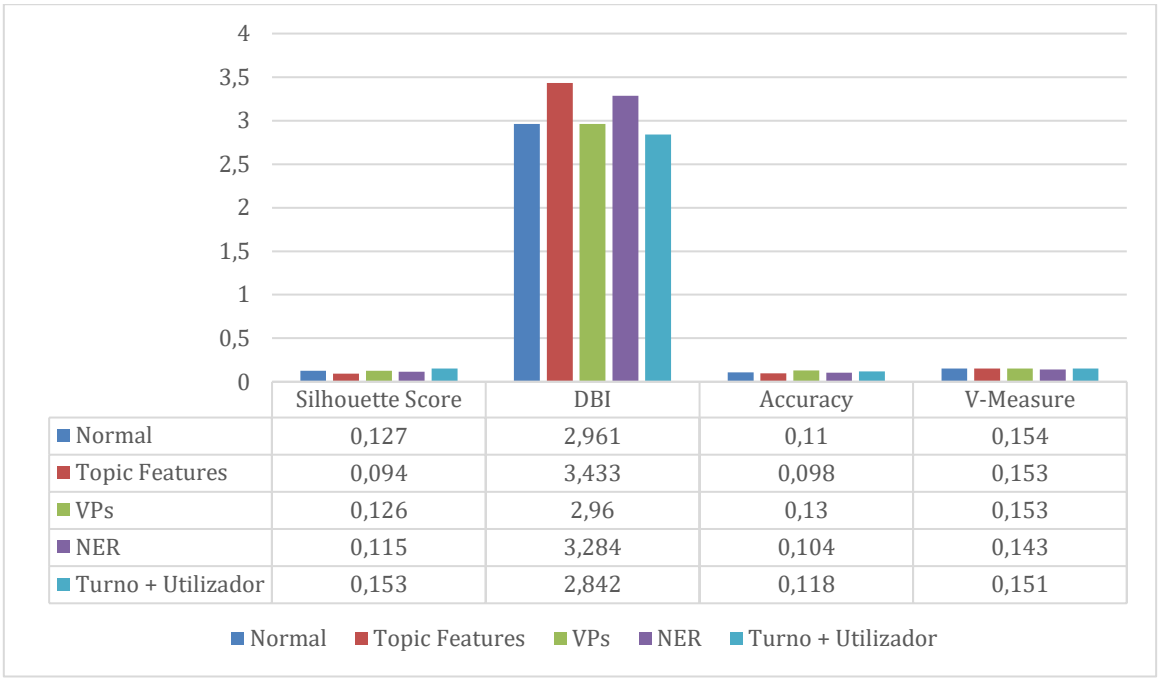
As utilizações das diversas combinações de parâmetros de TF-IDF não foram conclusivas, variando sempre conforme o método de avaliação ou corpora. Mais uma vez, o corpus CamRest676 é o que apresenta os melhores valores para a avaliação interna, tendo praticamente sempre valores acima de 0.1 para o *Silhouette Score*, algo que não acontece em mais nenhum corpus. E o SGD e MultiWOZ mantém-se com os melhores valores de avaliação externa, na qual se destaca o valor de *accuracy* de 0.504, no MultiWOZ com *max-df* = 0.85 e *min-df* = 10, ou seja, neste caso, mais de metade das previsões estão corretas.

Como não há nenhuma regra de qual a melhor combinação de parâmetros, utilizar-se-á *max-df* = 0.80 e *min-df* = 10, pois são os valores intermédios.

Agora, serão comparadas quatro variantes com o TF-IDF normal (azul-escuro nas figuras seguintes), usando sempre *max-df* = 0.80 e *min-df* = 10. Essas variações, são:

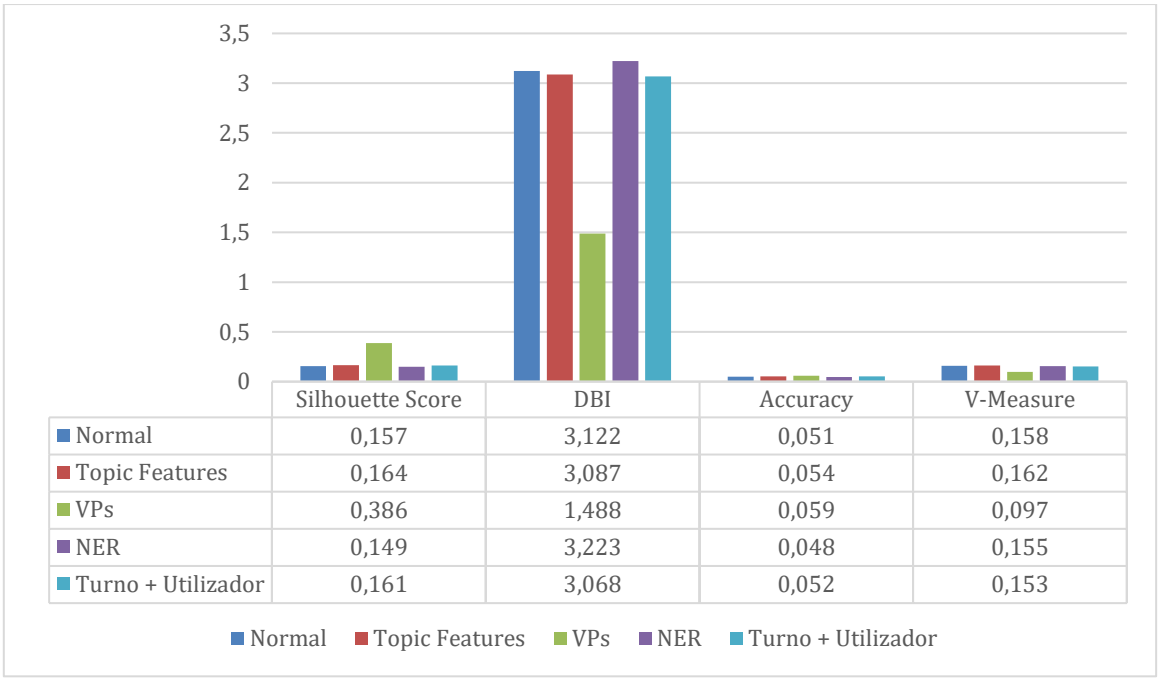
1. Uso de *Topic Modeling*. Remoção das 100 palavras mais relevantes em 10 tópicos, algo descoberto com o uso do LDA (*Topic Features* - vermelho).

2. Corpus transformado em VPs. Isto é, o corpus transformado em apenas sintagmas verbais (VPs - verde).
3. Corpus transformado com recurso ao NER, como referido no capítulo anterior (NER - roxo).
4. Inclusão de duas *features*. Uso de id de turno normalizado e o id de interlocutor. O objetivo é a tentativa de criar *features* relacionadas com o histórico de diálogo (Turno + Utilizador - azul-claro).



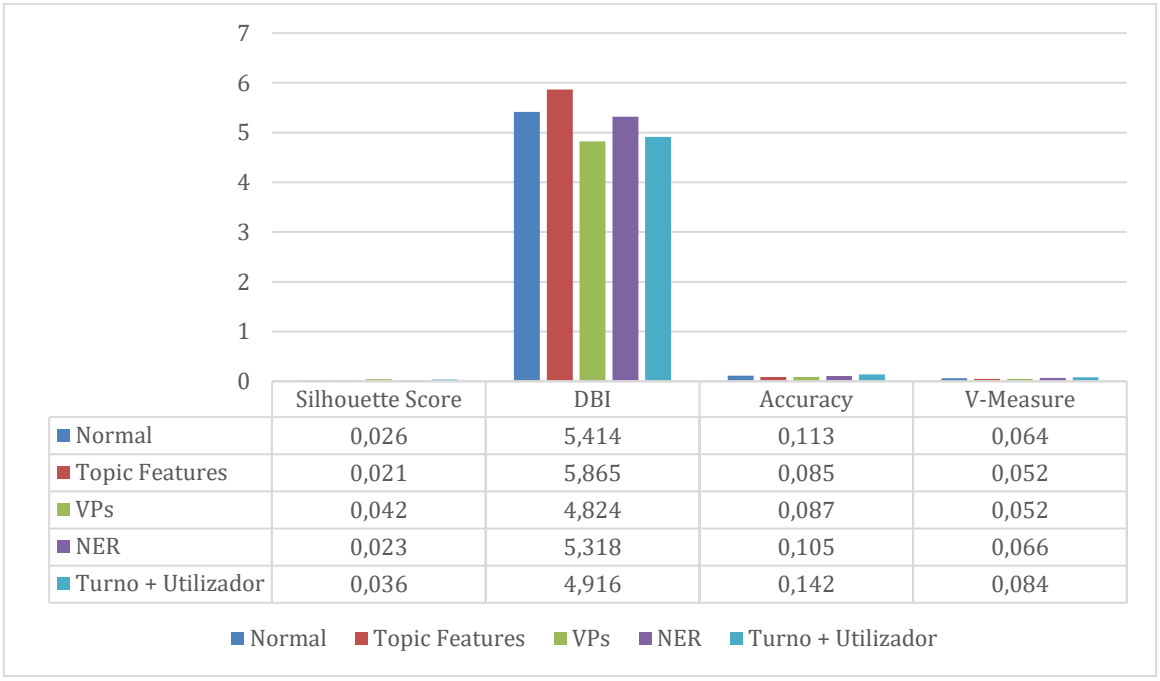
	Silhouette Score	DBI	Accuracy	V-Measure
Normal	0,127	2,961	0,11	0,154
Topic Features	0,094	3,433	0,098	0,153
VPs	0,126	2,96	0,13	0,153
NER	0,115	3,284	0,104	0,143
Turno + Utilizador	0,153	2,842	0,118	0,151

Figura 4.1 - Gráfico de colunas para testar a remoção das *features* de tópico, uso do corpus normal e o corpus composto apenas por sintagmas verbais aplicados ao CamRest676 de 4 atos de diálogo



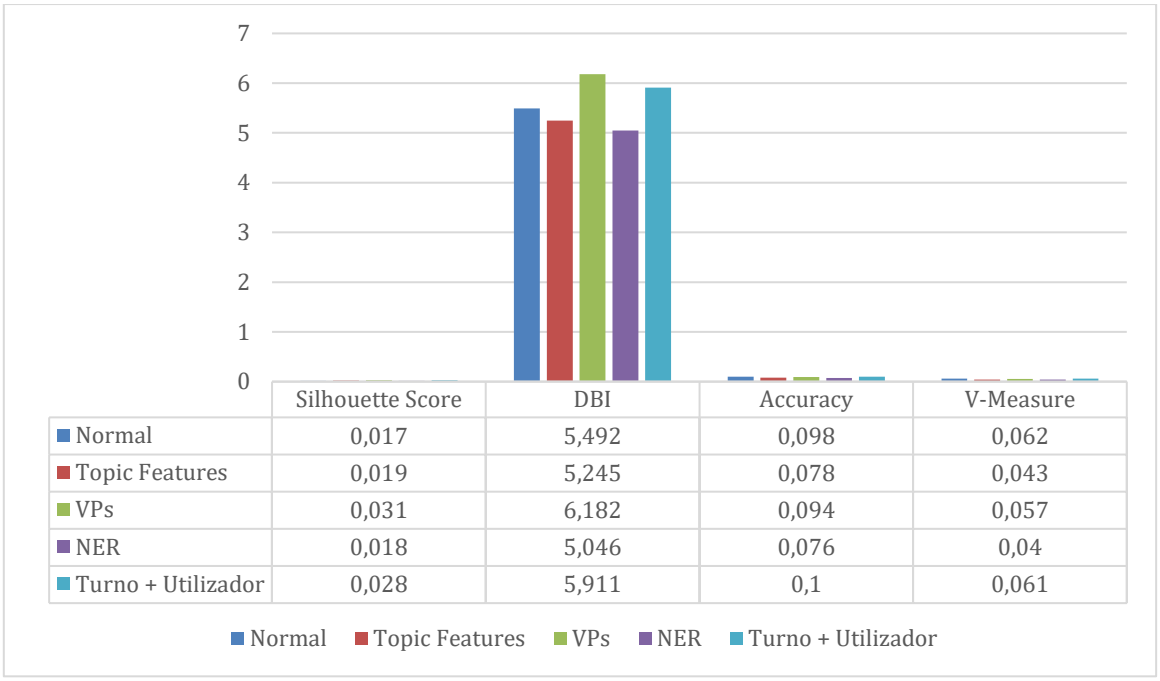
	Silhouette Score	DBI	Accuracy	V-Measure
Normal	0,157	3,122	0,051	0,158
Topic Features	0,164	3,087	0,054	0,162
VPs	0,386	1,488	0,059	0,097
NER	0,149	3,223	0,048	0,155
Turno + Utilizador	0,161	3,068	0,052	0,153

Figura 4.2 - Gráfico de colunas para testar a remoção das *features* de tópico, uso do corpus normal e o corpus composto apenas por sintagmas verbais aplicados ao CamRest676 de 16 atos de diálogo



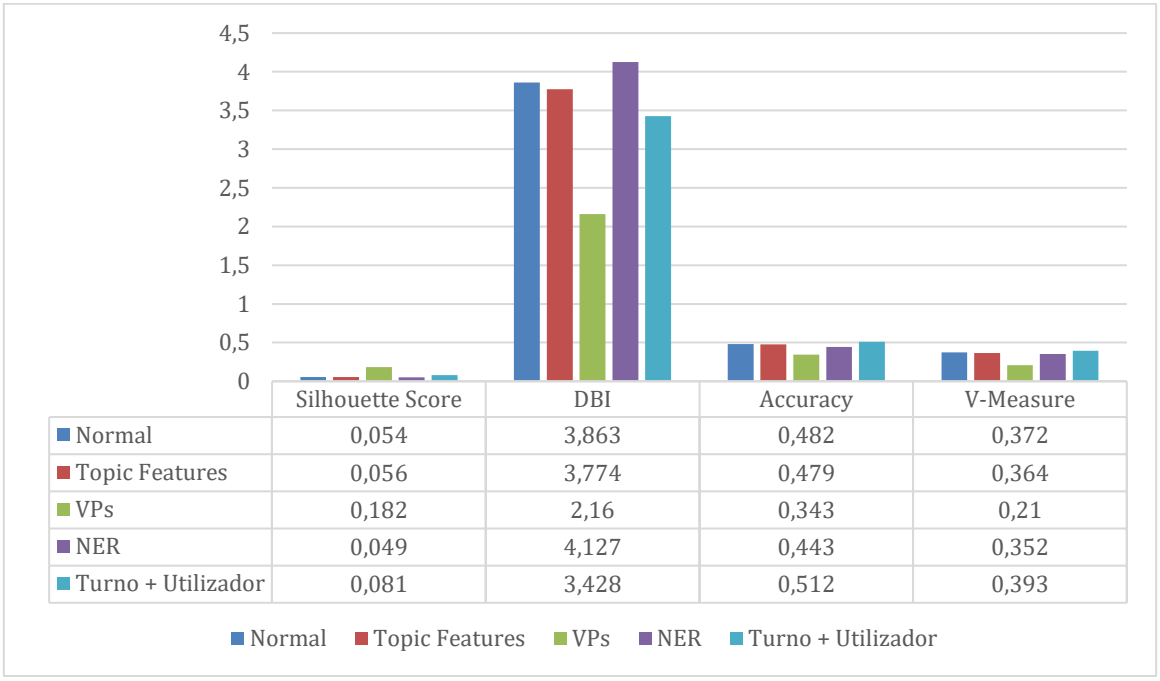
	Silhouette Score	DBI	Accuracy	V-Measure
Normal	0,026	5,414	0,113	0,064
Topic Features	0,021	5,865	0,085	0,052
VPs	0,042	4,824	0,087	0,052
NER	0,023	5,318	0,105	0,066
Turno + Utilizador	0,036	4,916	0,142	0,084

Figura 4.3 - Gráfico de colunas para testar a remoção das *features* de tópico, uso do corpus normal e o corpus composto apenas por sintagmas verbais aplicados ao Mastodon



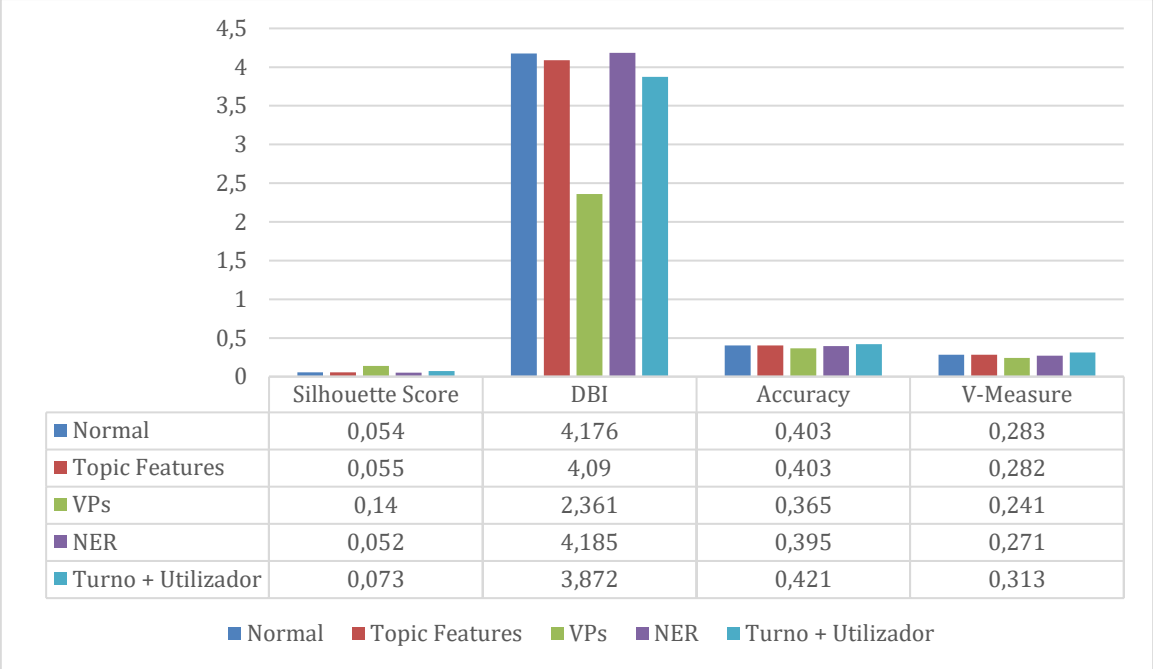
	Silhouette Score	DBI	Accuracy	V-Measure
Normal	0,017	5,492	0,098	0,062
Topic Features	0,019	5,245	0,078	0,043
VPs	0,031	6,182	0,094	0,057
NER	0,018	5,046	0,076	0,04
Turno + Utilizador	0,028	5,911	0,1	0,061

Figura 4.4 - Gráfico de colunas para testar a remoção das *features* de tópico, uso do corpus normal e o corpus composto apenas por sintagmas verbais aplicados ao DailyDialog



	Silhouette Score	DBI	Accuracy	V-Measure
Normal	0,054	3,863	0,482	0,372
Topic Features	0,056	3,774	0,479	0,364
VPs	0,182	2,16	0,343	0,21
NER	0,049	4,127	0,443	0,352
Turno + Utilizador	0,081	3,428	0,512	0,393

Figura 4.5 - Gráfico de colunas para testar a remoção das *features* de tópico, uso do corpus normal e o corpus composto apenas por sintagmas verbais aplicados ao MultiWOZ de atos de diálogo



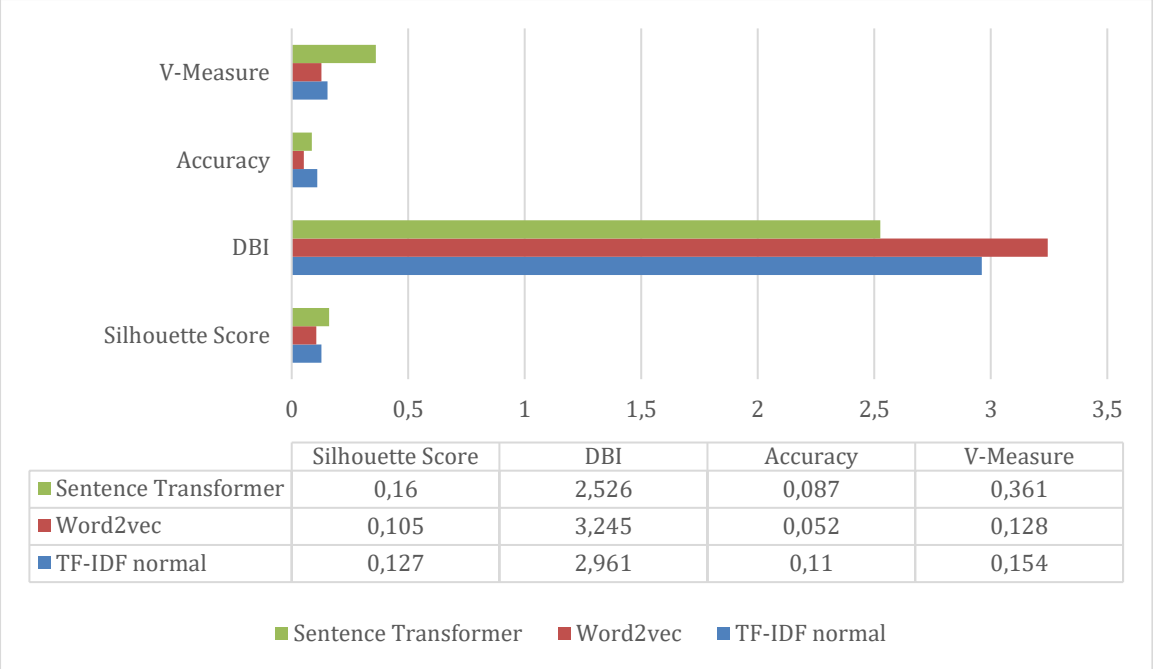
	Silhouette Score	DBI	Accuracy	V-Measure
Normal	0,054	4,176	0,403	0,283
Topic Features	0,055	4,09	0,403	0,282
VPs	0,14	2,361	0,365	0,241
NER	0,052	4,185	0,395	0,271
Turno + Utilizador	0,073	3,872	0,421	0,313

Figura 4.6 - Gráfico de colunas para testar a remoção das *features* de tópico, uso do corpus normal e o corpus composto apenas por sintagmas verbais aplicados ao SGD de atos de diálogo

Não há grandes técnicas em destaque. A utilização do TF-IDF normal com o uso das *features* de turno e utilizador melhorou quase sempre os resultados obtidos, mas ainda assim, nada significativo. A utilização do corpus em formato de VPs melhorou quase sempre a avaliação interna e piorou a externa, de notar que em certos casos o valor de *Silhouette Score* passou para o triplo, por exemplo no MultiWOZ.

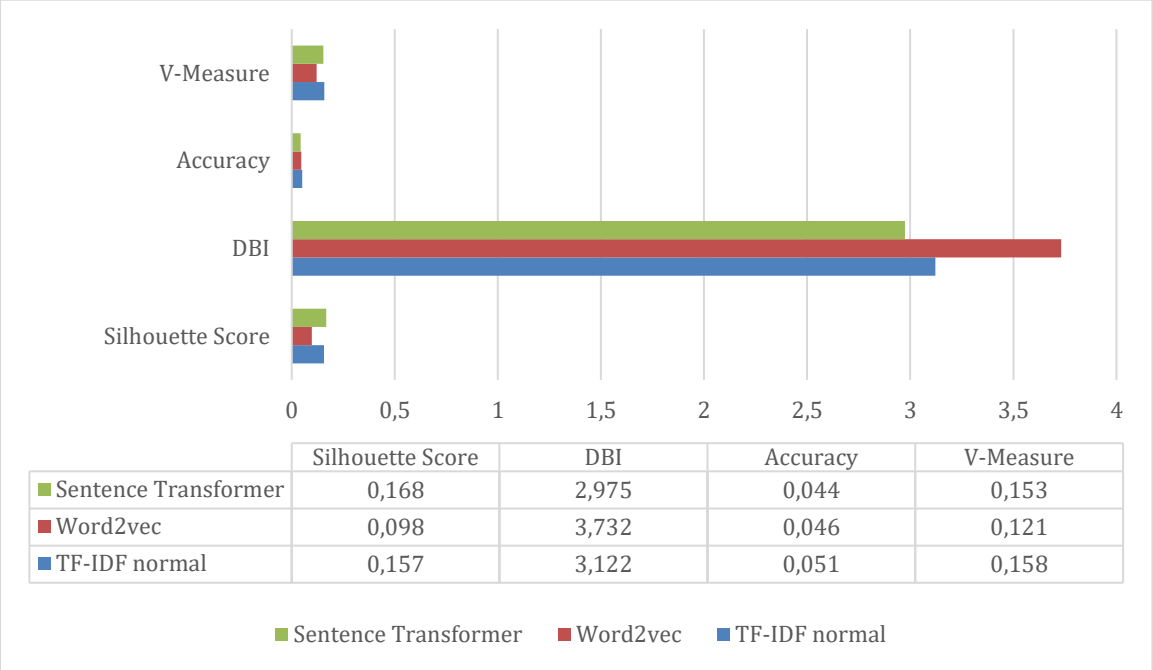
4.1.3 Modelos Word Embedding

Nesta seção serão comparadas as três abordagens de *Word Embedding*, e fazer uma avaliação de qual a melhor técnica.



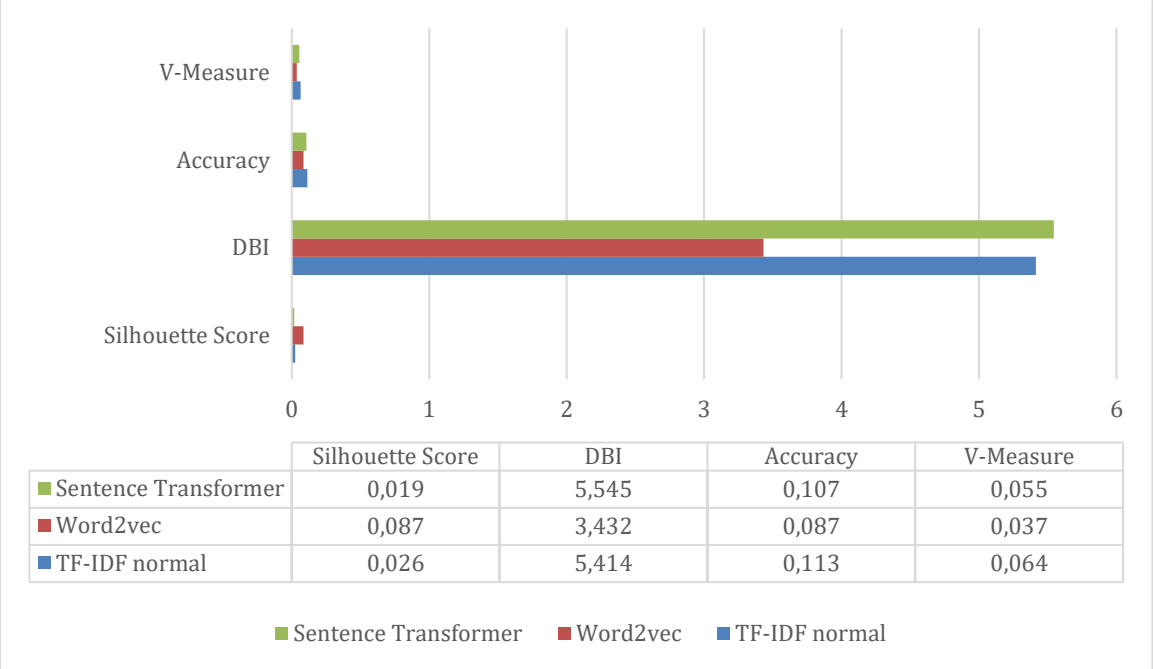
	Silhouette Score	DBI	Accuracy	V-Measure
Sentence Transformer	0,16	2,526	0,087	0,361
Word2vec	0,105	3,245	0,052	0,128
TF-IDF normal	0,127	2,961	0,11	0,154

Figura 4.7 - Comparação das três técnicas de *Word Embedding* para o CamRest676 de 4 atos de diálogo



	Silhouette Score	DBI	Accuracy	V-Measure
Sentence Transformer	0,168	2,975	0,044	0,153
Word2vec	0,098	3,732	0,046	0,121
TF-IDF normal	0,157	3,122	0,051	0,158

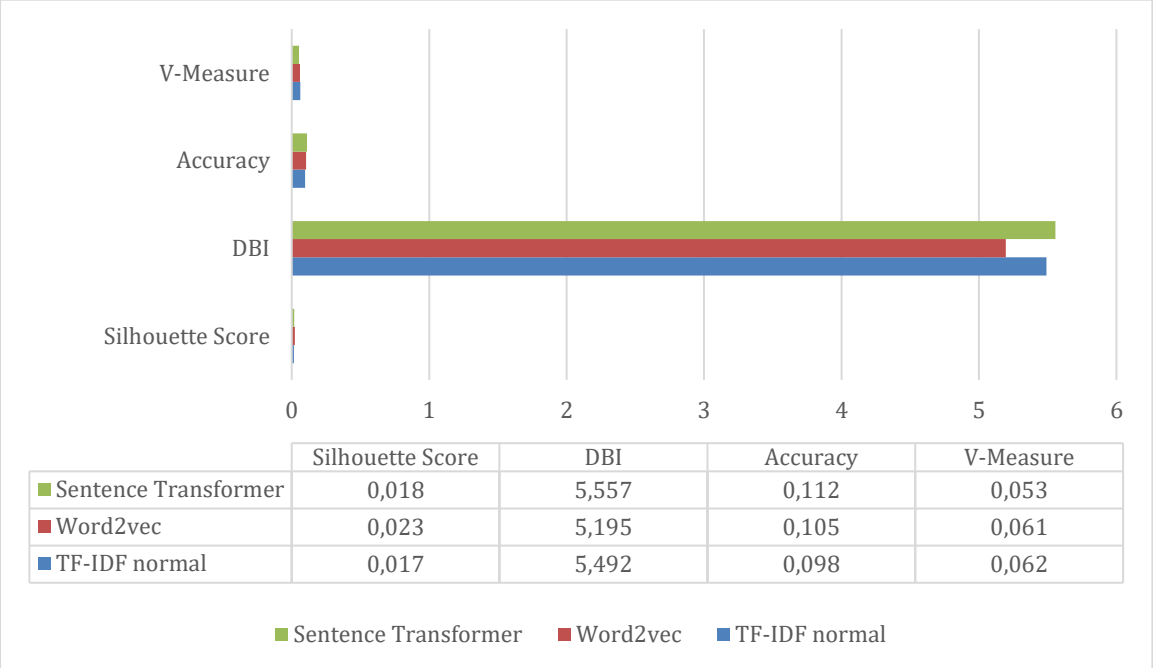
Figura 4.8 - Comparação das três técnicas de *Word Embedding* para o CamRest676 de 16 atos de diálogo



	Silhouette Score	DBI	Accuracy	V-Measure
Sentence Transformer	0,019	5,545	0,107	0,055
Word2vec	0,087	3,432	0,087	0,037
TF-IDF normal	0,026	5,414	0,113	0,064

Sentence Transformer Word2vec TF-IDF normal

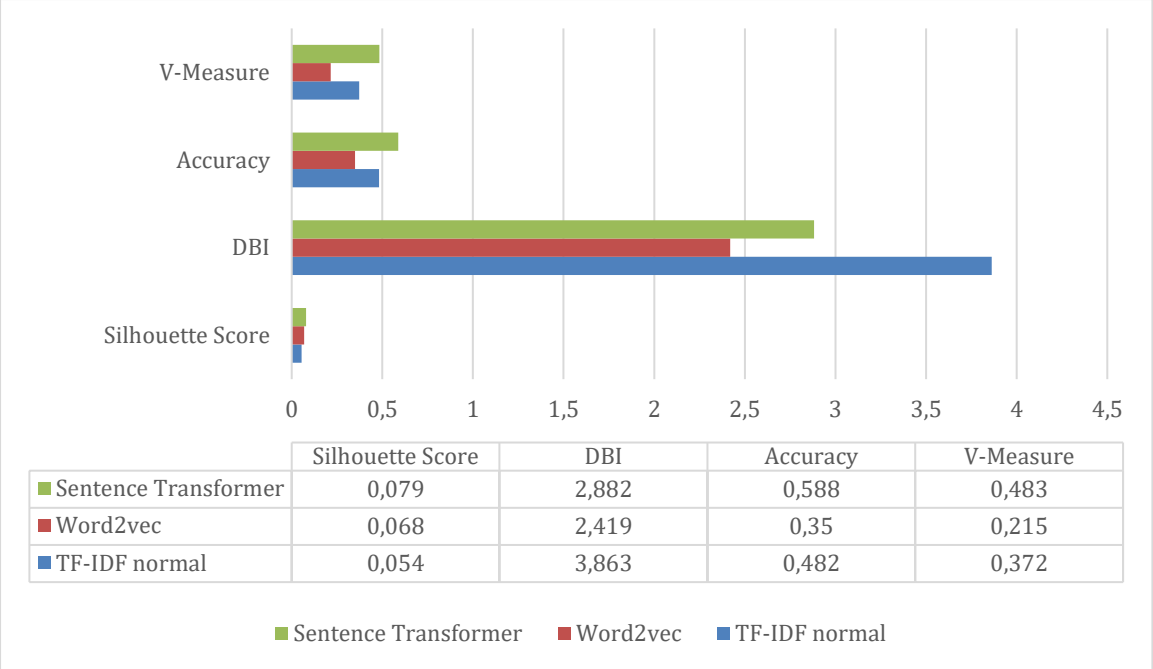
Figura 4.9 - Comparação das três técnicas de *Word Embedding* para o Mastodon



	Silhouette Score	DBI	Accuracy	V-Measure
Sentence Transformer	0,018	5,557	0,112	0,053
Word2vec	0,023	5,195	0,105	0,061
TF-IDF normal	0,017	5,492	0,098	0,062

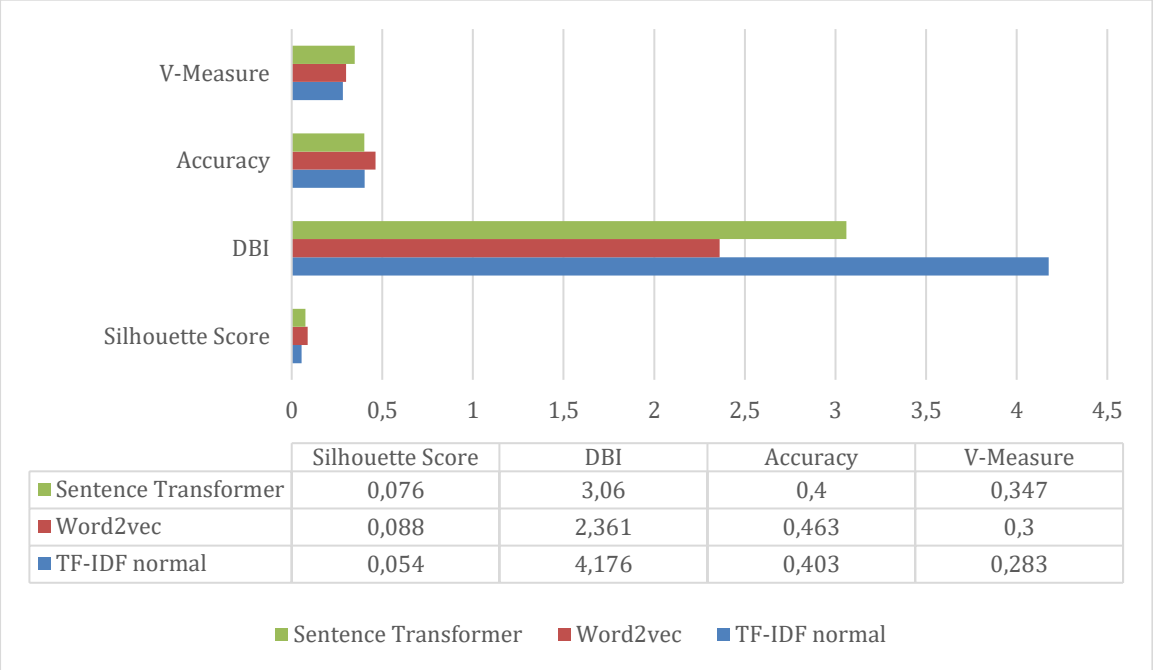
Sentence Transformer Word2vec TF-IDF normal

Figura 4.10 - Comparação das três técnicas de *Word Embedding* para o DailyDialog



	Silhouette Score	DBI	Accuracy	V-Measure
Sentence Transformer	0,079	2,882	0,588	0,483
Word2vec	0,068	2,419	0,35	0,215
TF-IDF normal	0,054	3,863	0,482	0,372

Figura 4.11 - Comparação das três técnicas de *Word Embedding* para o MultiWOZ de atos de diálogo



	Silhouette Score	DBI	Accuracy	V-Measure
Sentence Transformer	0,076	3,06	0,4	0,347
Word2vec	0,088	2,361	0,463	0,3
TF-IDF normal	0,054	4,176	0,403	0,283

Figura 4.12 - Comparação das três técnicas de *Word Embedding* para o SGD de atos de diálogo

A comparação entre as diferentes técnicas é quase unânime, o *Sentence Tranformer* é quem apresenta os melhores resultados para quase todas as métricas em quase todos os corpora, apesar de existir exceções. Por exemplo, para o SGD é o word2vec que tem os melhores valores, exceto na métrica V-Measure em que volta a ser o *Sentence*

Transformer. No CamRest para as duas variantes é o TF-IDF a apresentar os melhores valores na avaliação externa e o *Sentence Transformer* na avaliação interna.

4.1.4 Visualização dos Dados

São apresentados os gráficos PCA e t-SNE para os corpora CamRest676 de 4 e 16 *clusters*, Mastodon, DailyDialog, MultiWOZ e SGD, ambos com atos de diálogo nas suas *labels*. Estes gráficos foram obtidos recorrendo ao modelo de *Sentence Transformer*, o *all-MiniLM-L6-v2*.

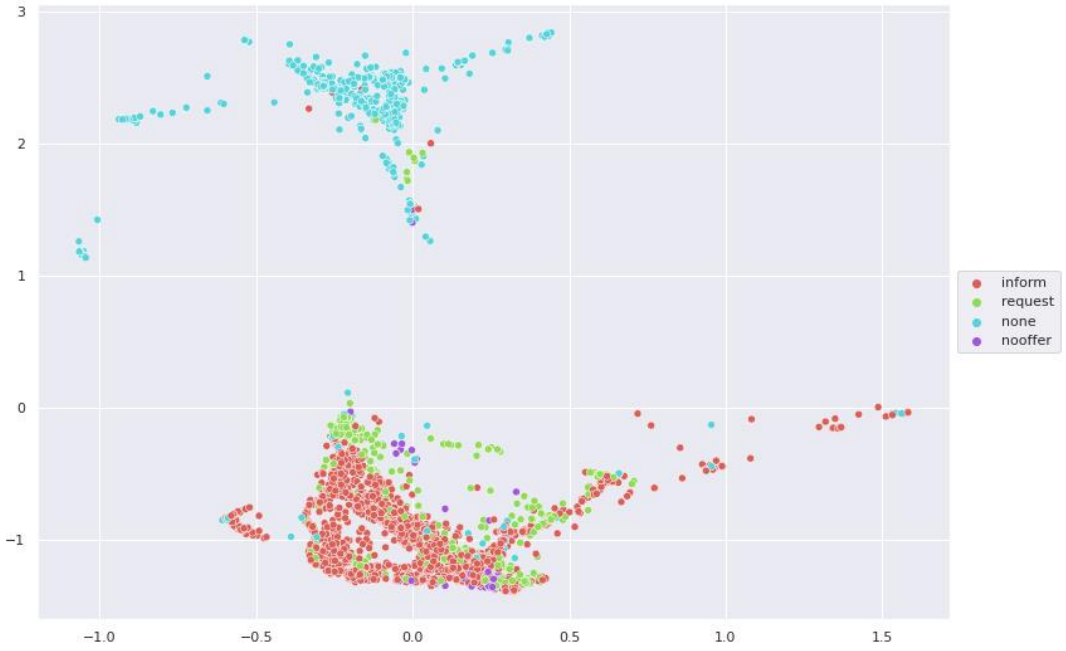


Figura 4.13 - t-SNE para o CamRest676 de 4 atos de diálogo

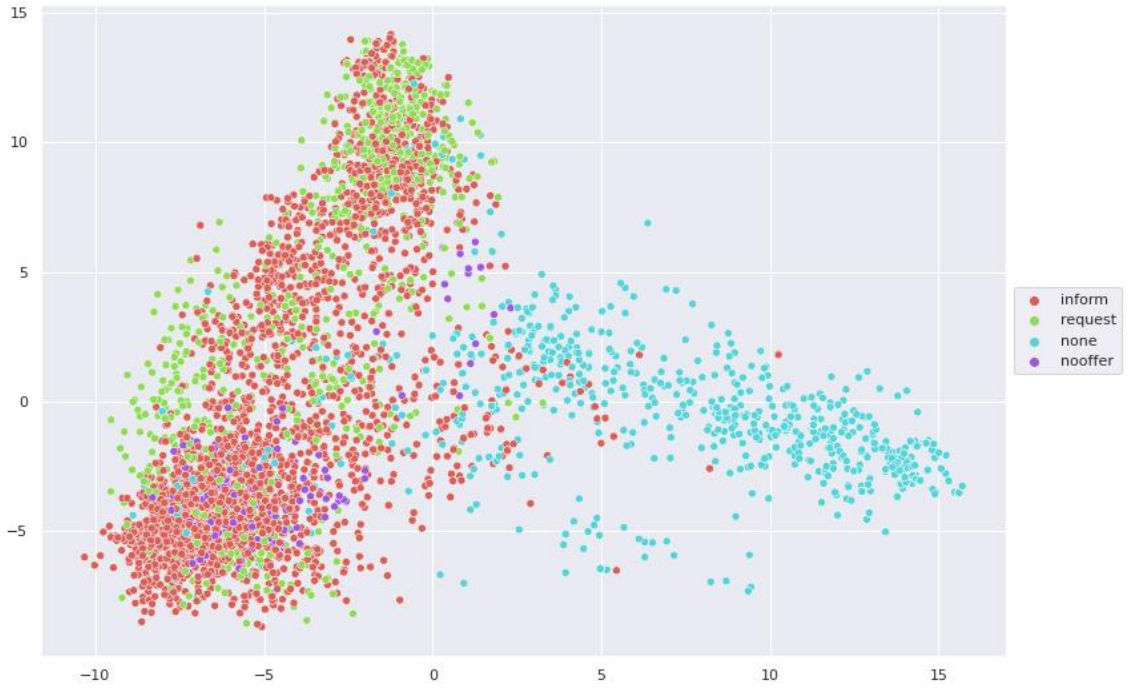


Figura 4.14 - PCA para o CamRest676 de 4 atos de diálogo

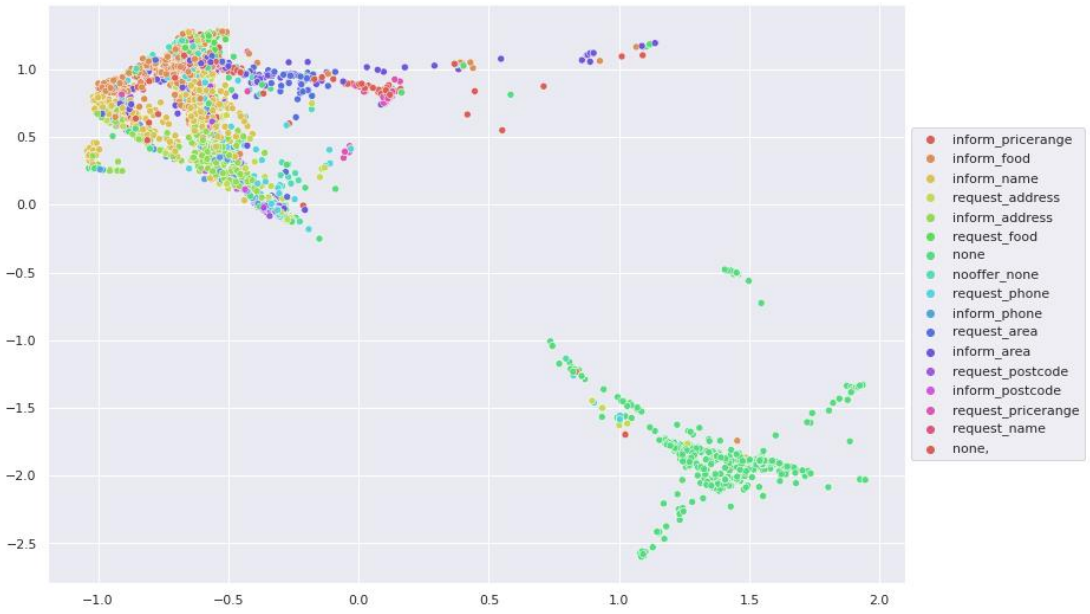


Figura 4.15 - t-SNE para o CamRest676 de 16 atos de diálogo

Figura 4.16 - PCA para o CamRest676 de 16 atos de diálogo

Estes quatro gráficos mostram bem a separação da *label "none"* comparado com as outras *labels*, o que justifica os resultados obtidos para a *V-Measure* neste corpus.

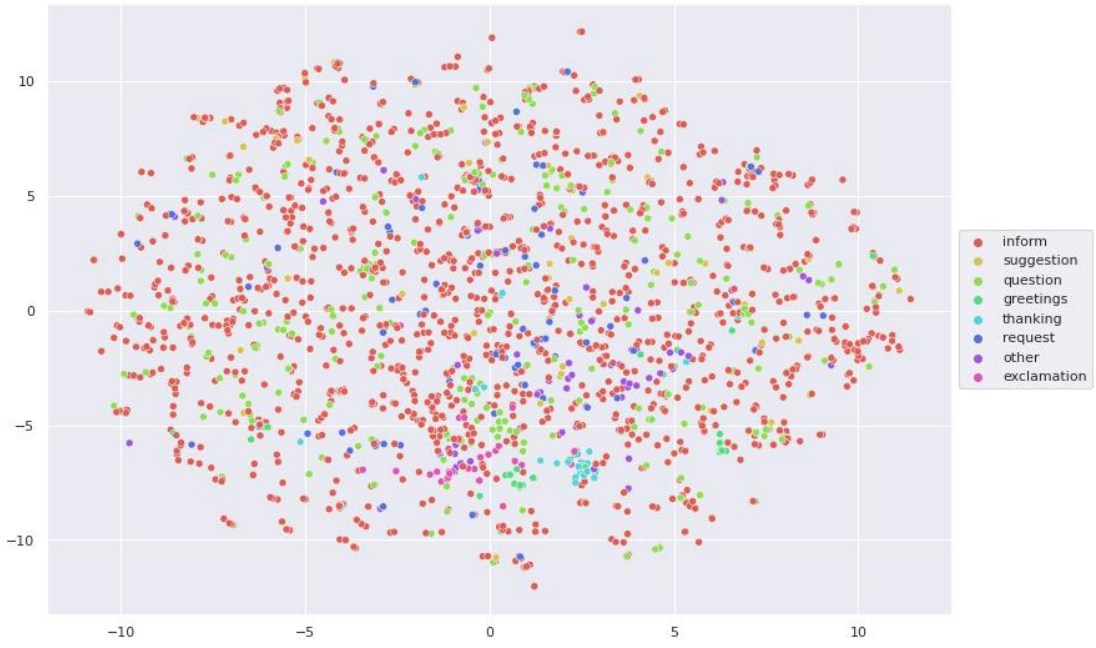


Figura 4.17 - t-SNE para o Mastodon

Figura 4.18 - PCA para o Mastodon

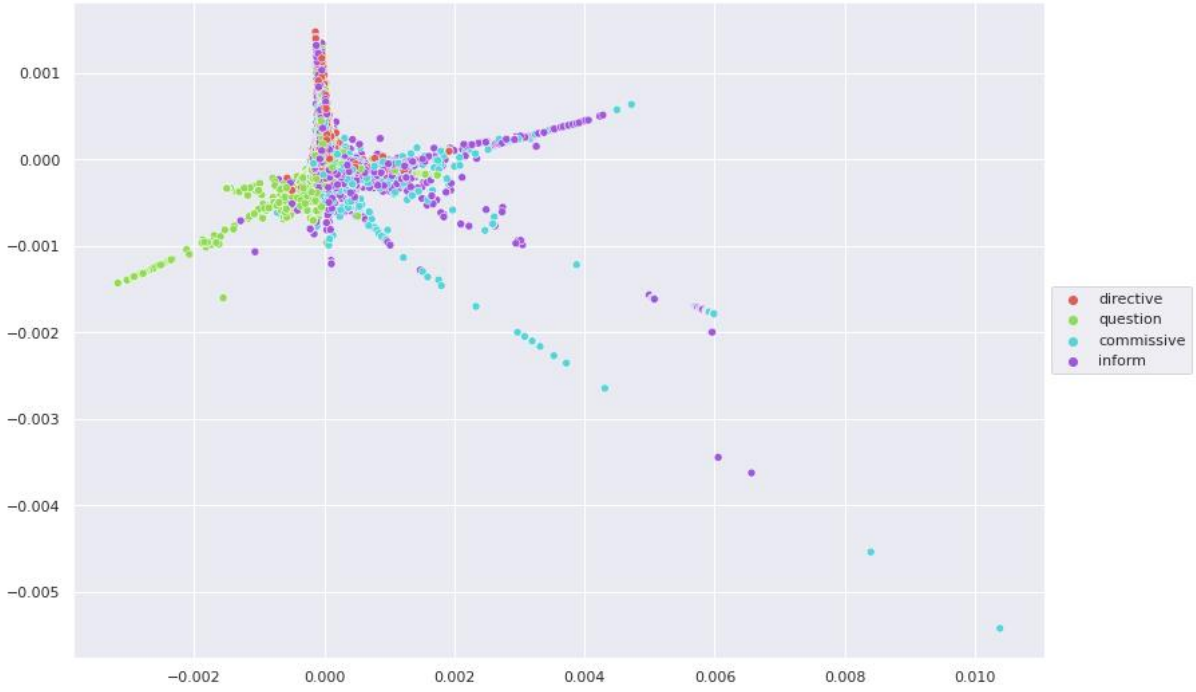


Figura 4.19 - t-SNE para o Daily Dialog

Figura 4.20 - PCA para o Daily Dialog

Os gráficos do DailyDialog e do Mastodon mostram uma confusão entre as diferentes classes, não existindo nenhuma separação entre elas, o que justifica os maus resultados obtidos anteriormente, comprovando que nestes corpora não foi possível chegar aos atos de diálogo. Ainda assim é possível observar na Figura 4.19 uma separação da classe "*question*" relativamente às outras.

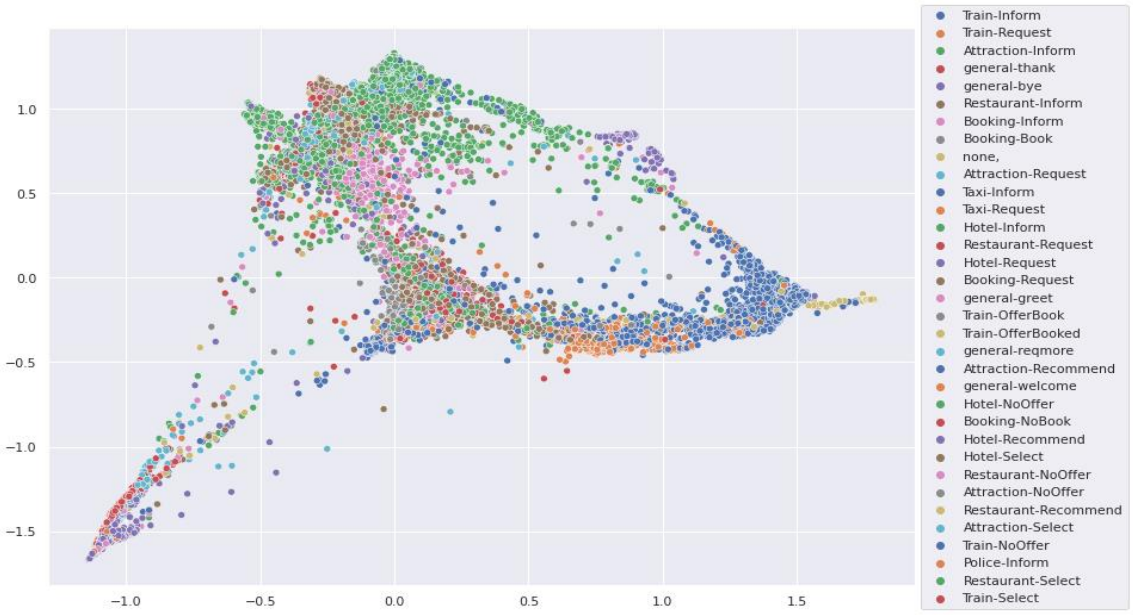


Figura 4.21 - t-SNE para o MultiWOZ de atos de diálogo

Figura 4.22 - PCA para o MultiWOZ de atos de diálogo

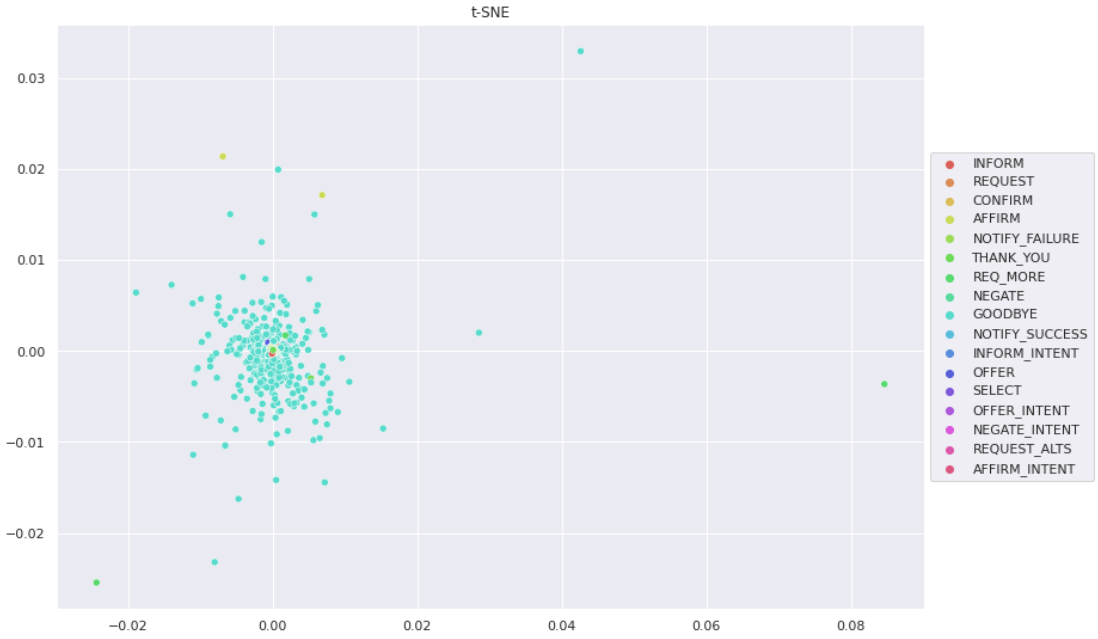


Figura 4.23 - t-SNE para o SGD de atos de diálogo

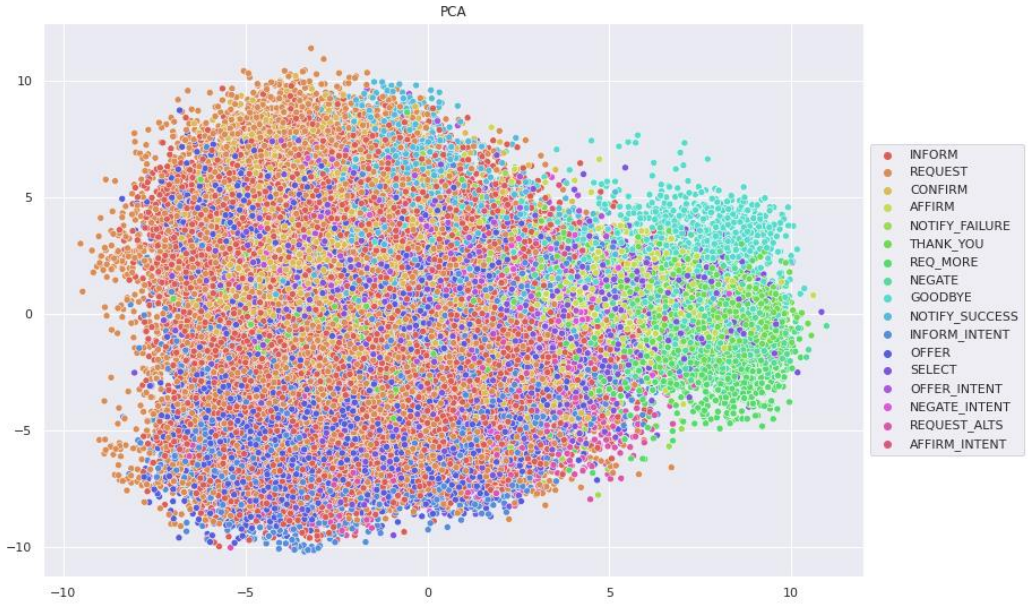


Figura 4.24 - PCA para o SGD de atos de diálogo

Estes dois corpora, o MultiWOZ e o SGD obtiveram resultados semelhantes e consideravelmente bons. NO PCA do SGD é possível ver a separação do *goodbye* (ciano), do *thank you* (verde) ao passo que outras *labels* se encontram mais misturadas. No MultiWOZ apesar de existirem muitas *labels*, o que dificulta os resultados, nota-se que elas acabam por estar juntas conforme o seu domínio, por exemplo hotel está agrupado mais abaixo no PCA.

4.2 Clustering para descobrir Intents

Neste subcapítulo serão descritas todas as experiências feitas na tentativa de agrupar os *intents* em *clusters*. Serão replicadas as mesmas abordagens do subcapítulo anterior, que foi na tentativa de descobrir atos de diálogo, este por sua vez, é na tentativa de descobrir *intents* através de abordagens não supervisionadas.

.

4.2.1 Modelos Sentence Transformer

Voltaram a ser testados os quatro modelos de *Sentence Transformer*, *all-mpnet-base-v2*, *all-distilroberta-v1*, *all-MiniLM-L12-v2* e *all-MiniLM-L6-v2*, de forma a analisar que modelo será utilizado futuramente.

Tabela 4.13 - Variações nos quatro modelos de *Sentence Transformer* aplicados ao SNIPS

Modelo *Sentence Transformer*	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
all-MiniLM-L6-v2	0.084	3.605	0.138	0.814
all-MiniLM-L12-v2	0.089	3.585	0.144	0.816
all-distilroberta-v1	0.101	3.191	0.161	**0.864**
all-mpnet-base-v2	**0.109**	**3.161**	**0.164**	0.859

Tabela 4.14 - Variações nos quatro modelos de *Sentence Transformer* aplicados ao MultiWOZ de *intents*

Modelo *Sentence Transformer*	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
all-MiniLM-L6-v2	**0.103**	2.981	0.462	0.489
all-MiniLM-L12-v2	**0.103**	**2.888**	0.457	0.488
all-distilroberta-v1	0.083	3.051	0.463	0.502
all-mpnet-base-v2	0.088	3.071	**0.475**	**0.514**

Tabela 4.15 - Variações nos quatro modelos de *Sentence Transformer* aplicados ao SGD de *intents*

Modelo *Sentence Transformer*	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
all-MiniLM-L6-v2	0.113	2.763	0.386	0.389
all-MiniLM-L12-v2	0.109	2.744	0.381	0.4
all-distilroberta-v1	0.105	2.819	0.398	0.396
all-mpnet-base-v2	0.1	2.789	0.393	0.401

Analisando as tabelas Tabela 4.13, Tabela 4.14 e Tabela 4.15, os resultados obtidos pelos quatro modelos são muito semelhantes, não justificando por isso a utilização de algum modelo mais robusto, por isso opta-se pelo modelo *all-MiniLM-L6-v2*, que de todos é o mais rápido. Esse será o modelo do *Sentence Transformer* que será comparado com as outras técnicas de *Word Embedding*. Nestes resultados é de destacar os elevados valores de *V-Measure* do SNIPS, sempre acima de 0.8. De destaque também a *V-Measure* nos outros dois corpora, sempre acima dos 0.35, inclusive atingindo no MultiWOZ os 0.5, bem como os bons valores de *accuracy* no SGD e MultiWOZ.

4.2.2 TF-IDF

Nesta seção serão feitos diversos testes para tentar obter o melhor proveito do TF-IDF, nomeadamente testes com diversas combinações de parâmetros, alterações ao conjunto de dados, na procura de conseguir o melhor resultado possível.

Assim como as experiências para os atos de diálogo, também foram testados diferentes valores dos parâmetros *max-df* e *min-df*. Para o *max-df* foram testados três valores: 0.75; 0.80; 0;85. Para o *min-df* foram testados três valores: 05; 10; 15.

Tabela 4.16 - Variações nos parâmetros do TF-IDF aplicados ao SNIPS

Parâmetros/Métricas	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
max-df:0.75; min-df:05	0.047	4.835	0.304	0.587
max-df:0.80; min-df:05	0.048	4.886	0.286	0.559
max-df:0.85; min-df:05	0.042	4.973	0.253	**0.613**
max-df:0.75; min-df:10	0.056	4.639	0.292	0.599
max-df:0.80; min-df:10	0.054	4.713	**0.314**	0.603
max-df:0.85; min-df:10	**0.057**	**4.459**	0.301	0.601
max-df:0.75; min-df:15	0.049	4.673	0.3	0.598
max-df:0.80; min-df:15	0.5	4.821	0.312	0.583
max-df:0.85; min-df:15	0.046	4.525	0.297	0.593

Tabela 4.17 - Variações nos parâmetros do TF-IDF aplicados ao MultiWOZ de *intents*

Parâmetros/Métricas	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
max-df:0.75; min-df:05	0.056	4.454	0.342	**0.336**
max-df:0.80; min-df:05	0.056	4.346	0.324	0.325
max-df:0.85; min-df:05	0.054	4.403	0.324	0.314
max-df:0.75; min-df:10	0.048	4.343	0.33	0.317
max-df:0.80; min-df:10	0.054	4.289	**0.335**	0.315
max-df:0.85; min-df:10	0.058	4.238	0.326	0.321
max-df:0.75; min-df:15	**0.063**	**4.15**	0.327	0.319
max-df:0.80; min-df:15	0.062	4.201	0.322	0.315
max-df:0.85; min-df:15	0.062	4.24	0.33	0.327

Tabela 4.18 - Variações nos parâmetros do TF-IDF aplicados ao SGD de *intents*

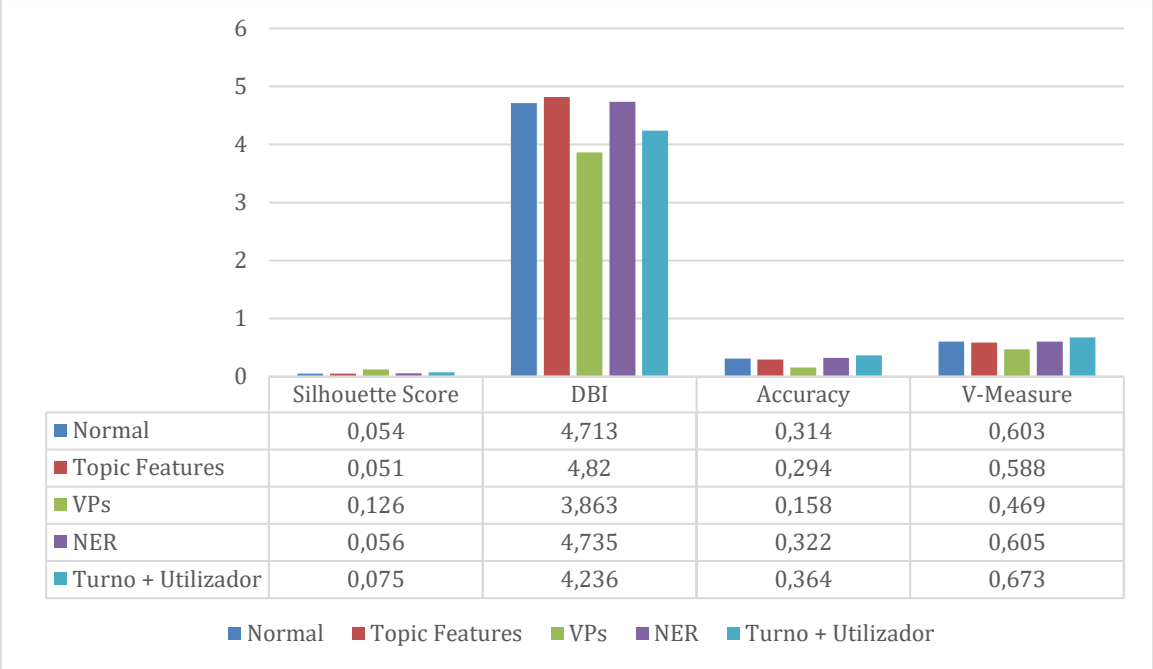
Parâmetros/Métricas	*Silhouette Score*	*DBI*	*Accuracy*	*V-Measure*
max-df:0.75; min-df:05	0.088	3.42	0.193	0.149
max-df:0.80; min-df:05	0.087	3.399	0.19	0.141
max-df:0.85; min-df:05	0.088	3.424	0.196	0.156
max-df:0.75; min-df:10	0.089	3.372	0.197	0.145
max-df:0.80; min-df:10	0.091	3.44	**0.206**	0.159
max-df:0.85; min-df:10	0.092	3.363	**0.206**	**0.161**
max-df:0.75; min-df:15	**0.096**	**3.329**	0.197	0.151
max-df:0.80; min-df:15	0.091	3.416	0.202	0.156
max-df:0.85; min-df:15	0.089	3.387	0.19	0.135

Mais uma vez, não houve conclusão definitiva sobre as melhores combinações de parâmetros para o TF-IDF.

Agora serão avaliadas e comparadas quatro vertentes estudadas no Estado da Arte:

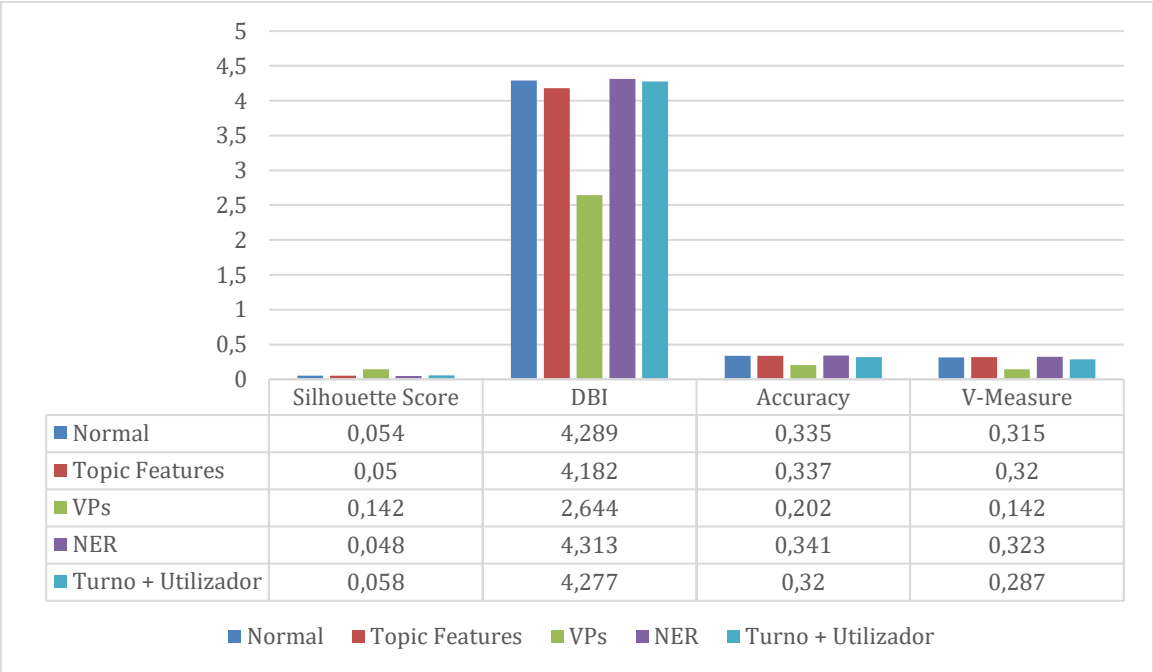
1. O impacto da remoção das palavras que indiciem tópicos, serão utilizadas 100 *Topic Features* (Topic Features);
2. A redução dos corpora na sua forma original aos sintagmas verbais (VPs);
3. Aplicação da técnica NER aos corpora (NER);
4. Uso do turno normalizado e do utilizador como *features* (Turno + Utilizador);

Estas quatro vertentes serão comparadas com o TF-IDF normal, como a sua variação dos parâmetros não apresentou nenhum resultado de destaque, e para manter a concordância serão utilizados os parâmetros *max-df* a 0.80 e *min-df* a 10.



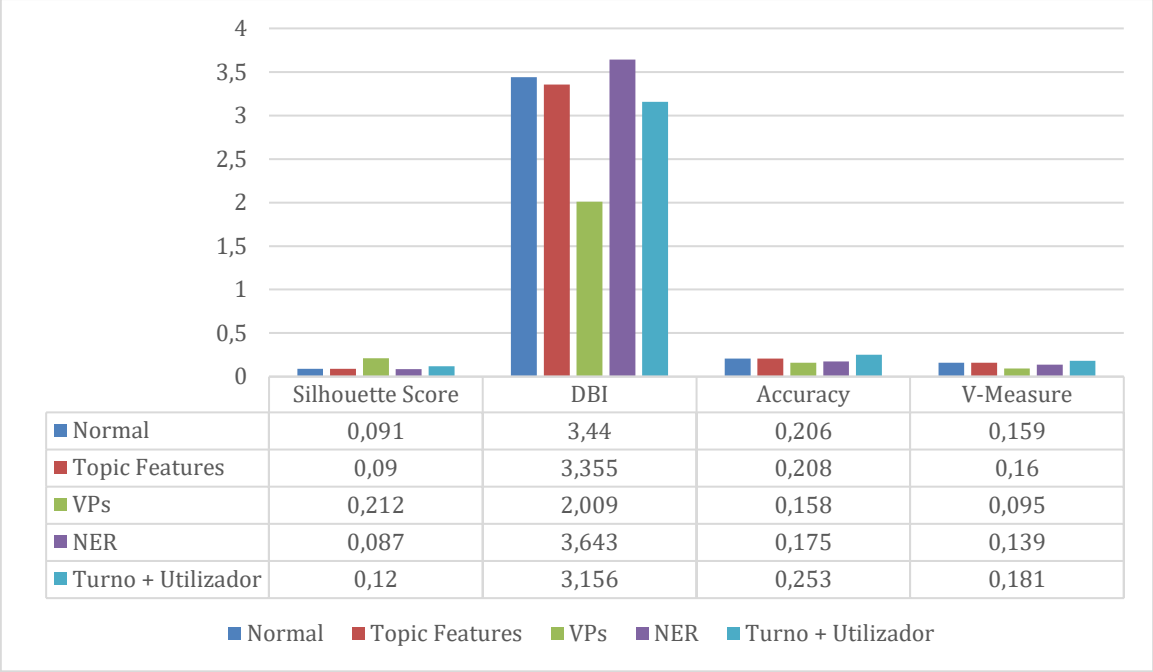
	Silhouette Score	DBI	Accuracy	V-Measure
Normal	0,054	4,713	0,314	0,603
Topic Features	0,051	4,82	0,294	0,588
VPs	0,126	3,863	0,158	0,469
NER	0,056	4,735	0,322	0,605
Turno + Utilizador	0,075	4,236	0,364	0,673

Figura 4.25 - Gráfico de colunas para testar a remoção das *features* de tópico, uso do corpus normal e o corpus composto apenas por sintagmas verbais aplicados ao SNIPS



	Silhouette Score	DBI	Accuracy	V-Measure
Normal	0,054	4,289	0,335	0,315
Topic Features	0,05	4,182	0,337	0,32
VPs	0,142	2,644	0,202	0,142
NER	0,048	4,313	0,341	0,323
Turno + Utilizador	0,058	4,277	0,32	0,287

Figura 4.26 - Gráfico de colunas para testar a remoção das *features* de tópico, uso do corpus normal e o corpus composto apenas por sintagmas verbais aplicados ao MultiWOZ de *intents*



	Silhouette Score	DBI	Accuracy	V-Measure
Normal	0,091	3,44	0,206	0,159
Topic Features	0,09	3,355	0,208	0,16
VPs	0,212	2,009	0,158	0,095
NER	0,087	3,643	0,175	0,139
Turno + Utilizador	0,12	3,156	0,253	0,181

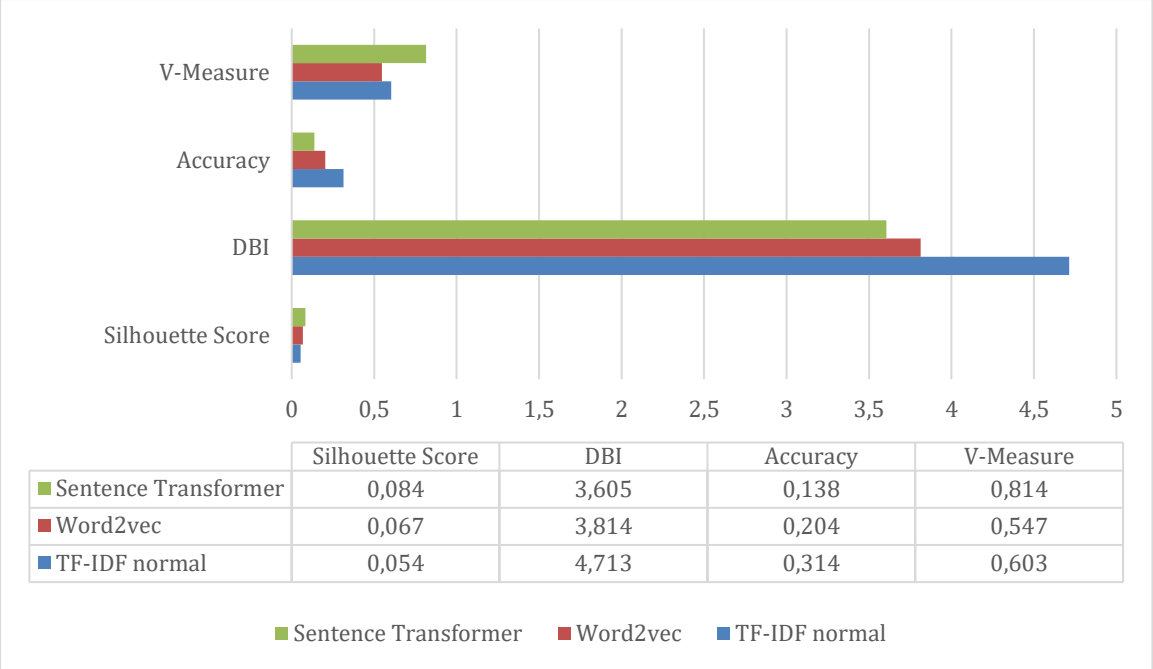
Figura 4.27 - Gráfico de colunas para testar a remoção das *features* de tópico, uso do corpus normal e o corpus composto apenas por sintagmas verbais aplicados ao SGD de *intents*

Como era pressuposto a utilização da técnica de remoção de *features* que indiciem determinado tópico não ajudou a melhorar os resultados obtidos anteriormente.

Os resultados obtidos também não são melhores que os anteriormente apresentados. No corpus SNIPS, a utilização de um corpus composto apenas por sintagmas verbais melhorou significativamente as métricas de avaliação interna e piorou por seu lado as métricas de avaliação externa, indicando que isso ajudou a uma melhor semelhança das instâncias pertencentes a cada *cluster*, mas ainda assim afastando-se do objetivo principal de agrupar em *intents*, o que piorou significativamente comparado com o corpus normal.

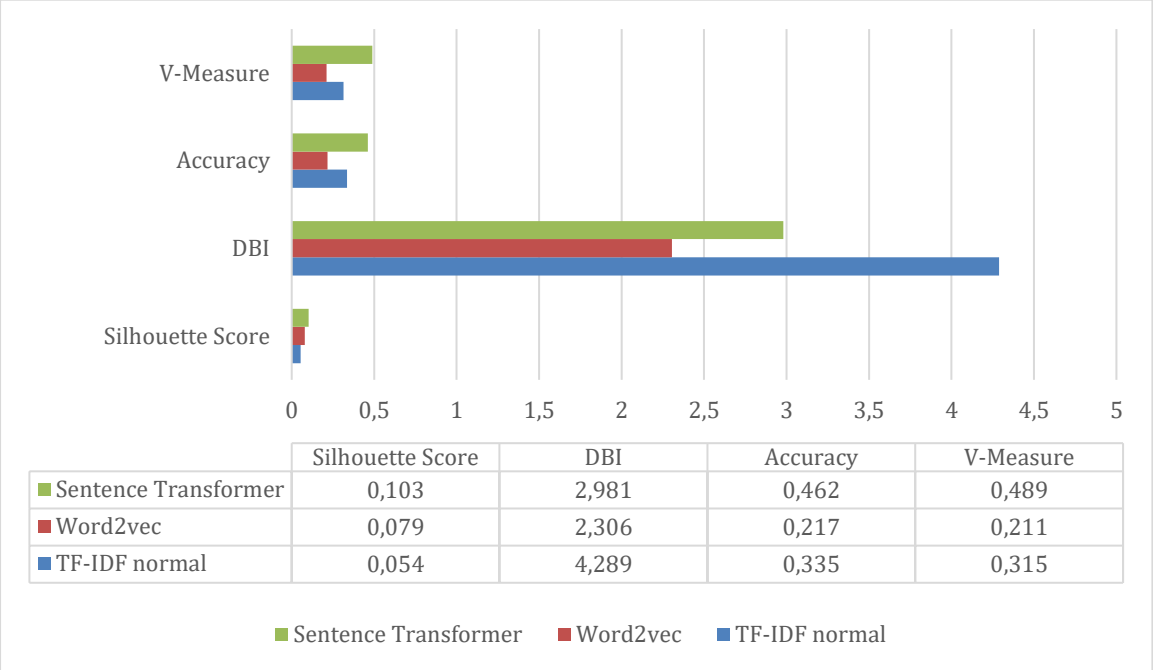
4.2.3 Modelos Word Embedding

Neste subcapítulo serão comparadas as três abordagens de *Word Embedding* utilizadas. Para ser coerente com os outros testes realizados neste capítulo, o TF-IDF será com as configurações normais (uso normal do corpus, sem novas *features*) com os parâmetros de *minDF* = 10 e o *maxDF* = 0.8. No *Sentence Transformer* será utilizado o modelo *all-MiniLM-L6*-v2, que apesar de ser o menos robusto dos quatro modelos utilizados, não teve resultados significativamente abaixo dos outros modelos. Estas duas abordagens serão comparadas com o word2vec que até então não tinha sido testado.



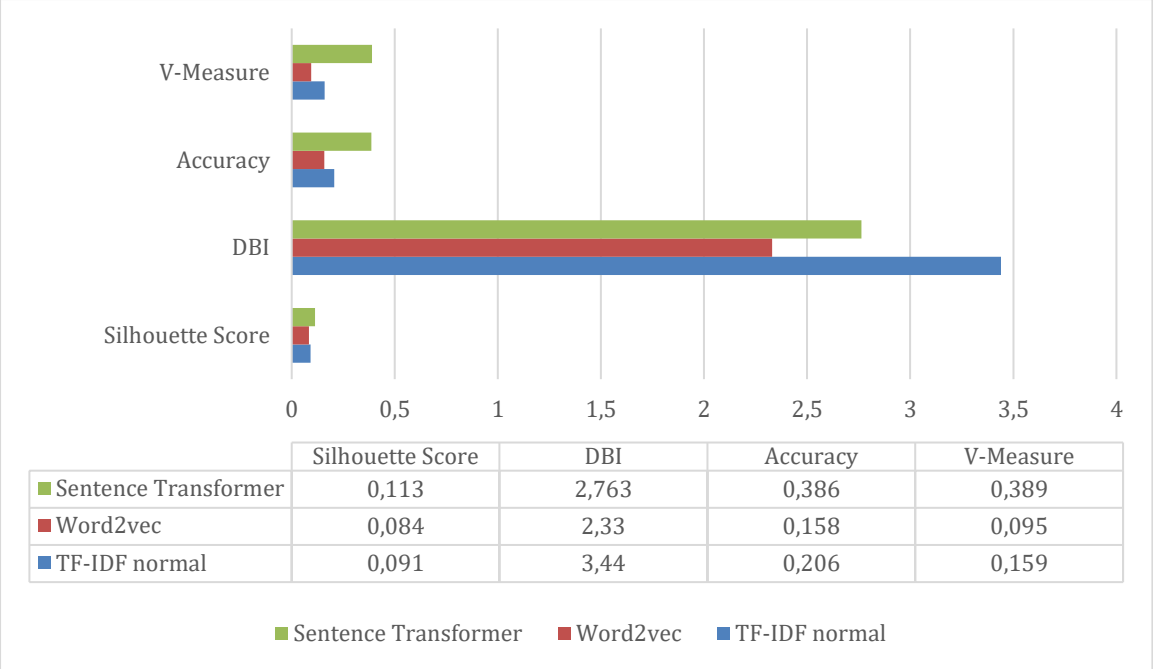
	Silhouette Score	DBI	Accuracy	V-Measure
Sentence Transformer	0,084	3,605	0,138	0,814
Word2vec	0,067	3,814	0,204	0,547
TF-IDF normal	0,054	4,713	0,314	0,603

Figura 4.28 - Comparação das três técnicas de Word Embedding para o SNIPS



	Silhouette Score	DBI	Accuracy	V-Measure
Sentence Transformer	0,103	2,981	0,462	0,489
Word2vec	0,079	2,306	0,217	0,211
TF-IDF normal	0,054	4,289	0,335	0,315

Figura 4.29 - Comparação das três técnicas de Word Embedding para o MultiWOZ de *intents*



	Silhouette Score	DBI	Accuracy	V-Measure
Sentence Transformer	0,113	2,763	0,386	0,389
Word2vec	0,084	2,33	0,158	0,095
TF-IDF normal	0,091	3,44	0,206	0,159

Figura 4.30 - Comparação das três técnicas de Word Embedding para o SGD de *intents*

O *Sentence Transformer* é o que obtém de longe os melhores valores comparativamente com as outras duas técnicas. O word2vec e TF-IDF apresentam resultados equiparáveis nos diferentes corpora realizados.

4.2.4 Visualização dos Dados

Serão apresentados os gráficos PCA e t-SNE para os corpora SNIPS, MultiWOZ e SGD, ambos com *intents* nas suas *labels*. Estes gráficos foram obtidos recorrendo ao modelo de *Sentence Transformer*, o *all-MiniLM-L6-v2*.

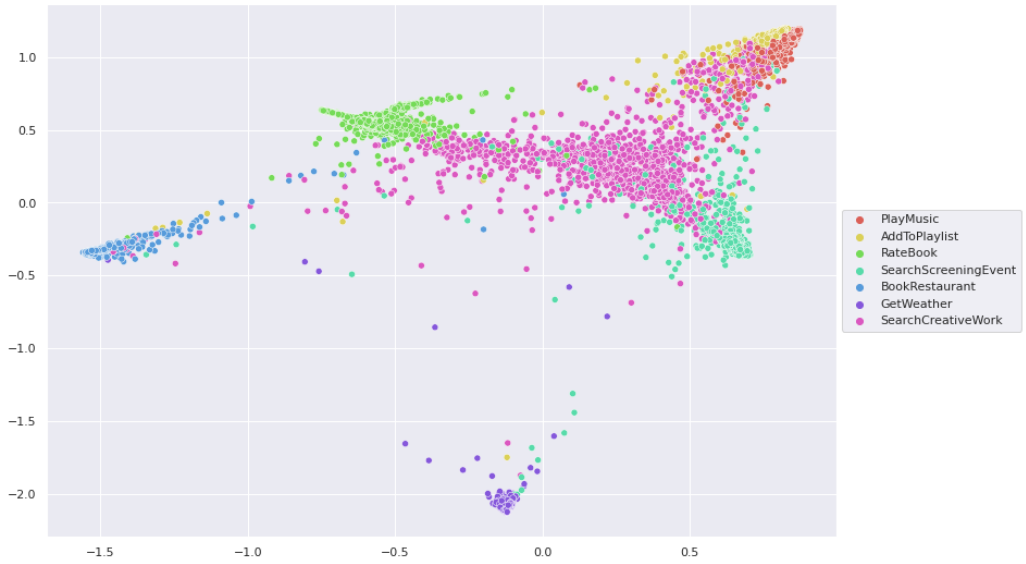


Figura 4.31 - t-SNE para o SNIPS

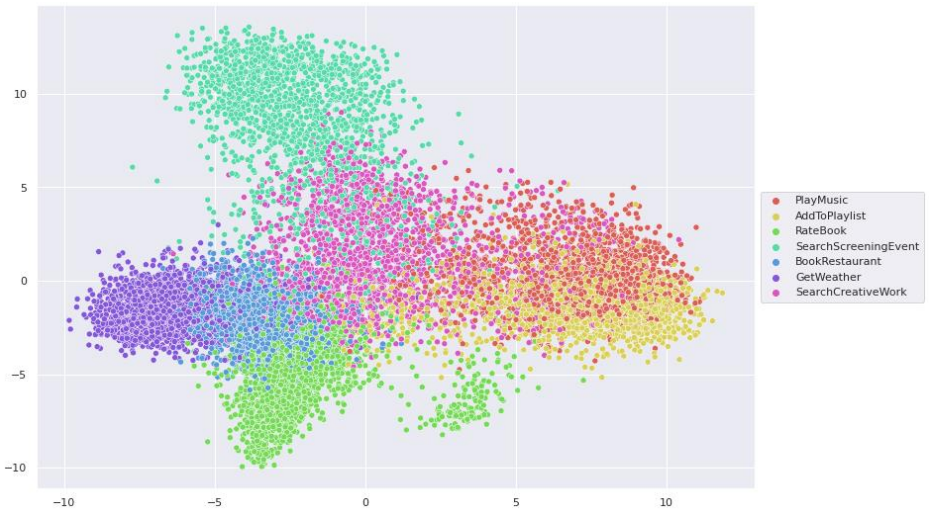


Figura 4.32 - PCA para o SNIPS

No PCA do SNIPS é visível a separação das cores, apesar de existir algumas labels que se encontram misturadas, nomeadamente *"Play Music"* a vermelho e *"AddToPlaylist"* a amarelo. Esta separação das cores justifica os bons resultados obtidos na métrica de *V-Measure*, ou seja, as instâncias de cada *cluster* são semelhantes e cada instância pertence ao *cluster* correto, os *clusters* foram bem definidos o que se comprova pelo PCA principalmente. Valores mais altos não foram obtidos talvez pelas *labels "Play Music"* e *"AddToPlaylist"* serem semelhantes, pois estão relacionadas com música.

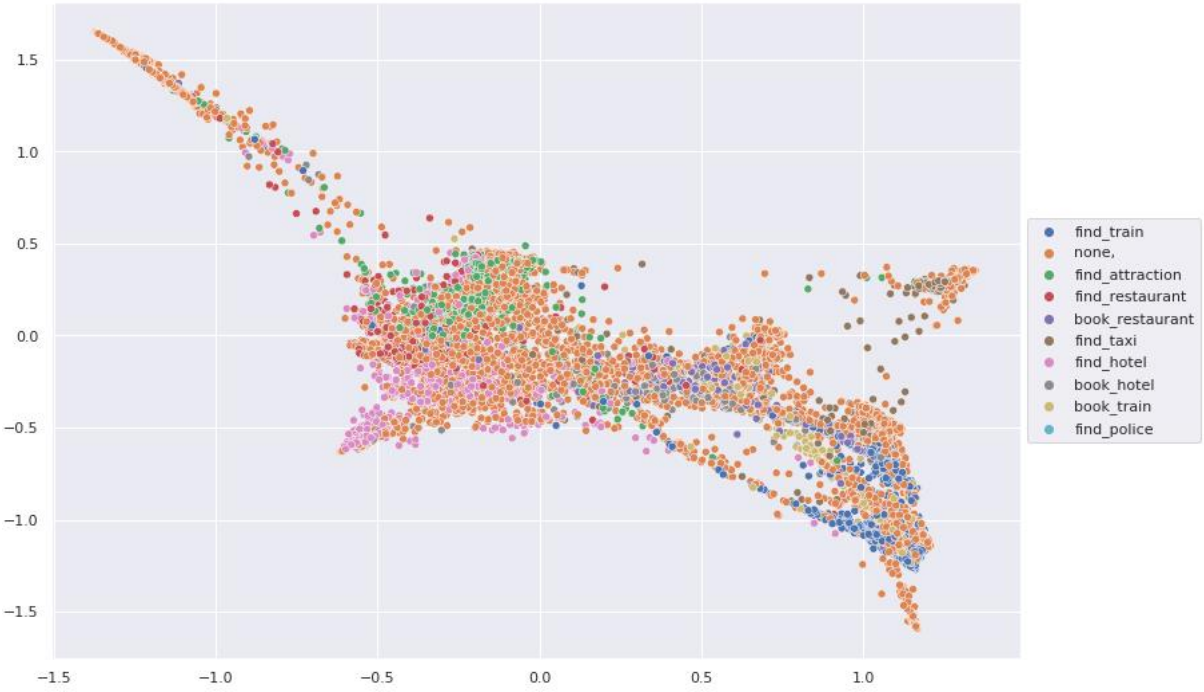


Figura 4.33 - t-SNE para o MultiWOZ de *intents*

Figura 4.34 - PCA para o MultiWOZ de *intents*

É possível observar o distanciamento das diferentes *labels*, o que mostra os bons resultados obtidos no MultiWOZ. Verifica-se por exemplo no PCA que as instâncias a azul (*find_train*) estão bem separadas das a verde (*find_attraction*), e ambas separadas das de vermelho (*find_restaurant*), por exemplo.

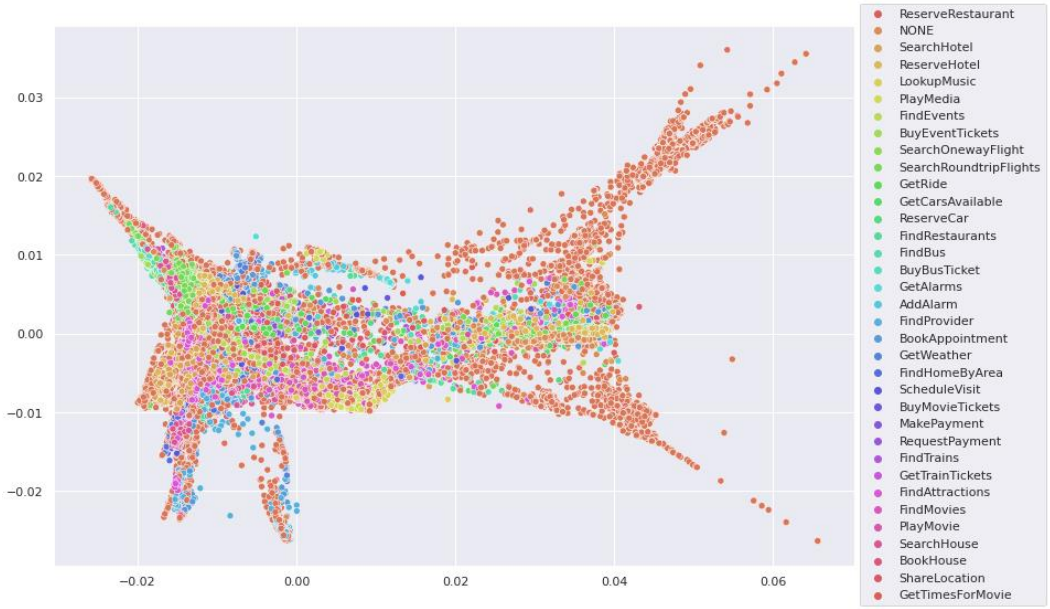


Figura 4.35 - t-SNE para o SGD de *intents*

Figura 4.36 - PCA para o SGD de atos de *intents*

No t-SNE do SGD é visível a separação de algumas das *labels*, ainda que outras estejam misturadas, o que até justifica os bons resultados obtidos na avaliação externa.

4.3 Considerações adicionais

Após análise dos resultados deste capítulo destacam-se algumas notas:

1. Os resultados não são muito conclusivos de que tipo de *label* as abordagens utilizadas se aproximam.
2. Nos *intents*, os resultados foram mais consistentes do que nos atos de diálogo.
3. Corpus como o DailyDialog e Mastodon estão longe de atingir os atos de diálogo. Isto pode ser justificado pelo facto do DailyDialog ter em grande predominância um ato de diálogo, o *inform*, e pelo facto do Mastodon ser de conversas de uma rede social, o que traz muito ruído.
4. Das técnicas utilizadas, destaca-se a redução dos corpora na sua forma original aos sintagmas verbais, tendo obtido melhores valores na avaliação interna, justificável pelo facto dos conjuntos de dados serem reduzidos e por isso haver maior semelhança entre as falas.
5. Das três técnicas de representação de falas é o *Sentence Transformer* que obtém os melhores resultados.

5 DESCOBERTA DE FLUXOS

Após o *clustering,* é descrito neste capítulo a etiquetagem dos respetivos *clusters* e a criação dos seus fluxos, que irão dar origem à orientação de um agente inteligente para conduzir uma conversa.

5.1 Etiquetagem de Clusters e Fluxos de Diálogo

Foram aplicadas as diferentes técnicas de etiquetagem descritas nos capítulos anteriores. As tabelas presentes no Anexo E - Tabelas com a aplicação das diferentes técnicas de etiquetagem a todos os corpora mostram as três formas de etiquetagem referidas anteriormente para cada um dos corpora utilizados neste trabalho.

Essas tabelas são exemplos de resultados ocorridos após uma iteração. Como já referido, o K-Means é não-determinístico, ou seja, há uma atualização dos centroides e a reatribuição das instâncias ao *cluster* mais próximo em cada iteração. Os fluxos obtidos e apresentados foram de acordo com as tabelas mencionadas.

Agora serão apresentados os respetivos fluxos com base nas tabelas do subcapítulo anterior, onde serão testadas as diferentes técnicas de etiquetagem para dar nome aos nós do grafo.

Os primeiros cinco fluxos apresentados nas figuras Figura 5.1, Figura 5.2, Figura 5.3, Figura 5.4 e Figura 5.5 são exemplos meramente ilustrativos baseados nas técnicas de etiquetagem anteriormente referidas. São ilustrativos, pois como ambos os corpora não apresentam uma *feature* relativa ao ID de diálogo torna-se impossível criar o *Start of Dialogue* e o *End of Dialogue*.

Os restantes fluxos terão duas novidades em relação aos anteriores. Será feita uma divisão entre Sistema e Utilizador e será um fluxo completo (de acordo com o número de *labels* originais) em conjunto com o SOD (*Start of Dialogue*) e o EOD (*End of Dialogue*). Isto será aplicado no SGD e no MultiWOZ, apenas para atos de diálogo e *intents*, respetivamente.

Para o corpus CamRest676 com 4 *clusters* e utilizando a técnica da fala mais próxima do centroide como modo de representação, obteve-se o fluxo presente na Figura 5.1.

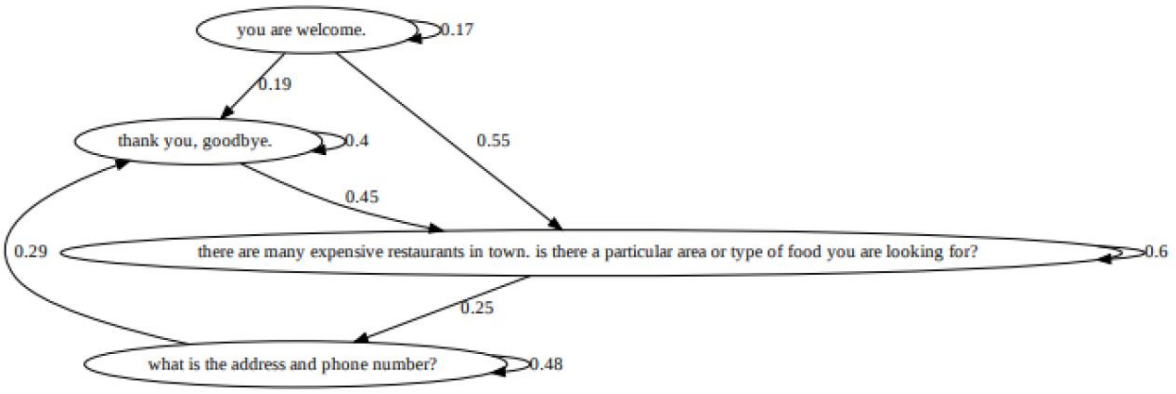


Figura 5.1 - Fluxo de diálogo CamRest676 com 4 *clusters*, com fala mais próxima do centroide

Para o corpus CamRest676 com 16 *clusters* e utilizando a técnica de predicado como modo de representação, obteve-se o fluxo presente na Figura 5.2.

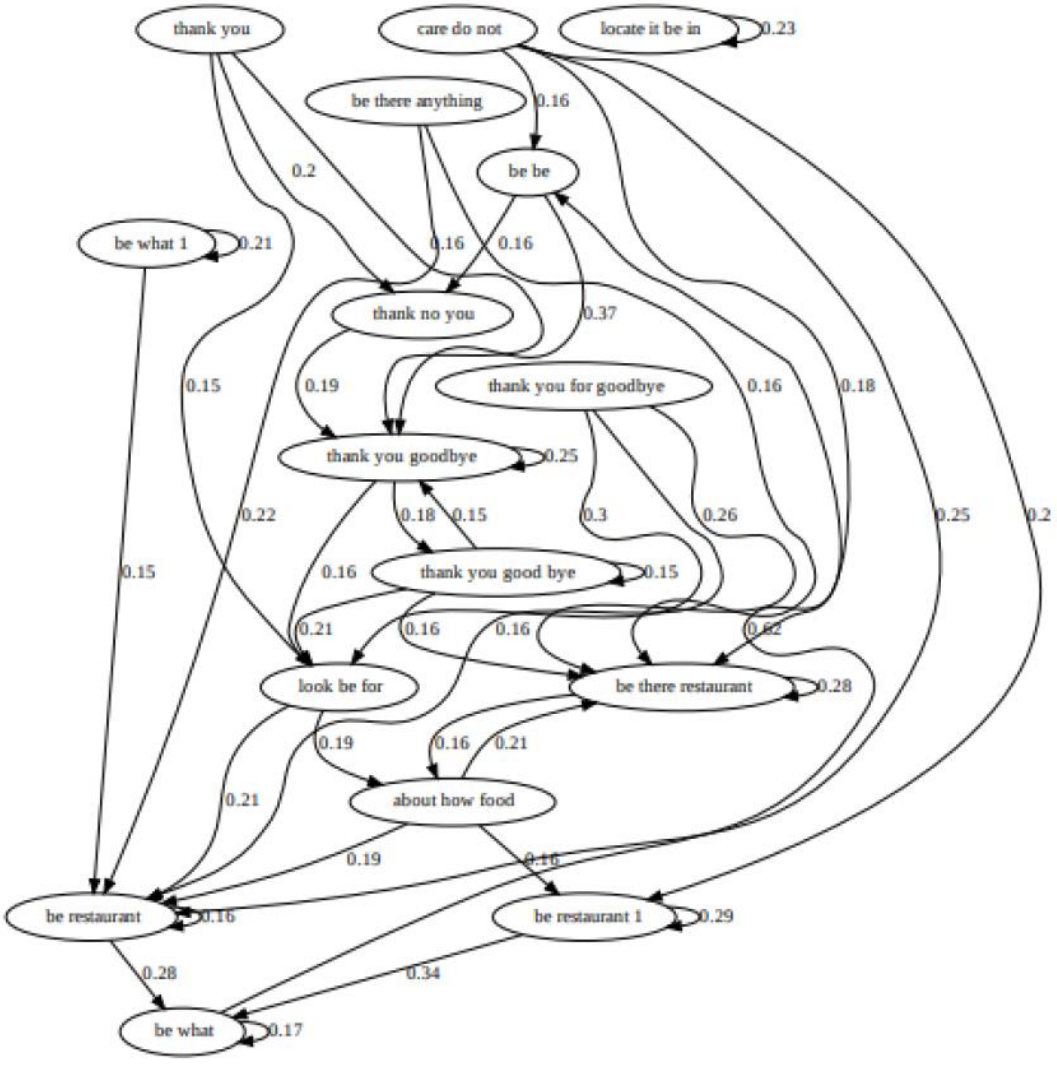


Figura 5.2 - Fluxo de diálogo CamRest676 com 16 *clusters*, com predicado

Para o corpus Mastodon e utilizando a técnica de Bigrama como modo de representação, obteve-se o fluxo presente na Figura 5.3.

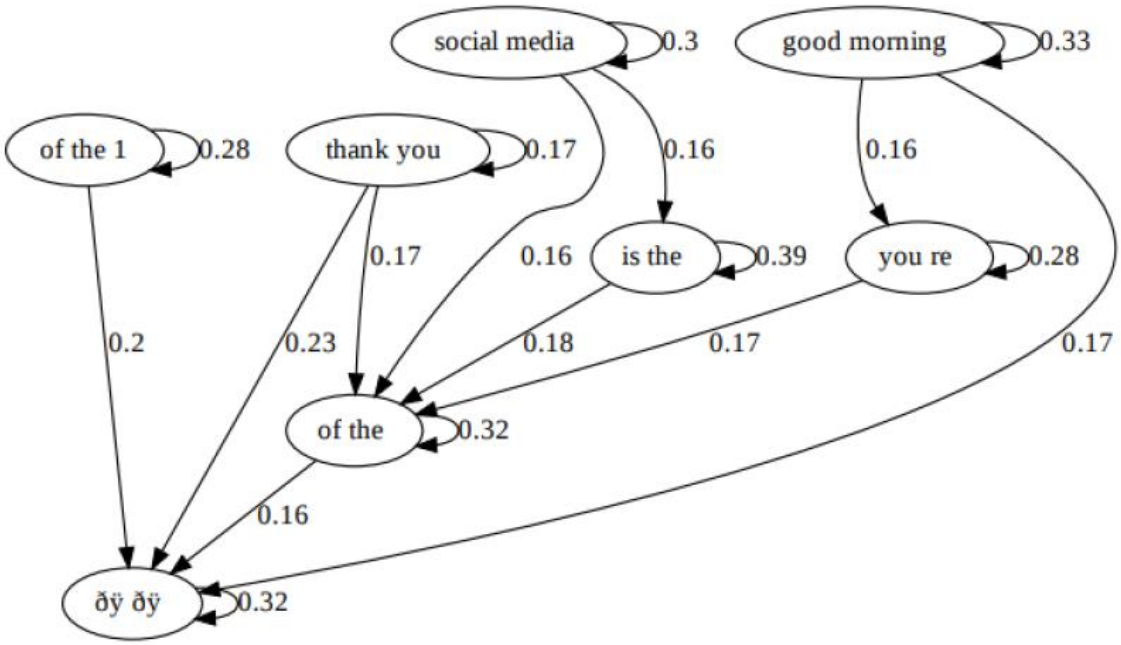


Figura 5.3 - Fluxo de diálogo Mastodon, com Bigrama

Para o corpus DailyDialog e utilizando a técnica da fala mais próxima do centroide como modo de representação, obteve-se o fluxo presente na Figura 5.4

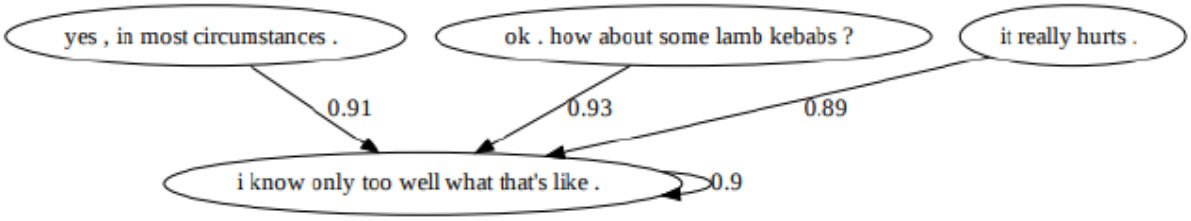


Figura 5.4 - Fluxo de diálogo DailyDialog, com Frase mais próxima do centroide

Para o corpus SNIPS e utilizando a técnica de Bigrama como modo de representação, obteve-se o fluxo presente na Figura 5.5.

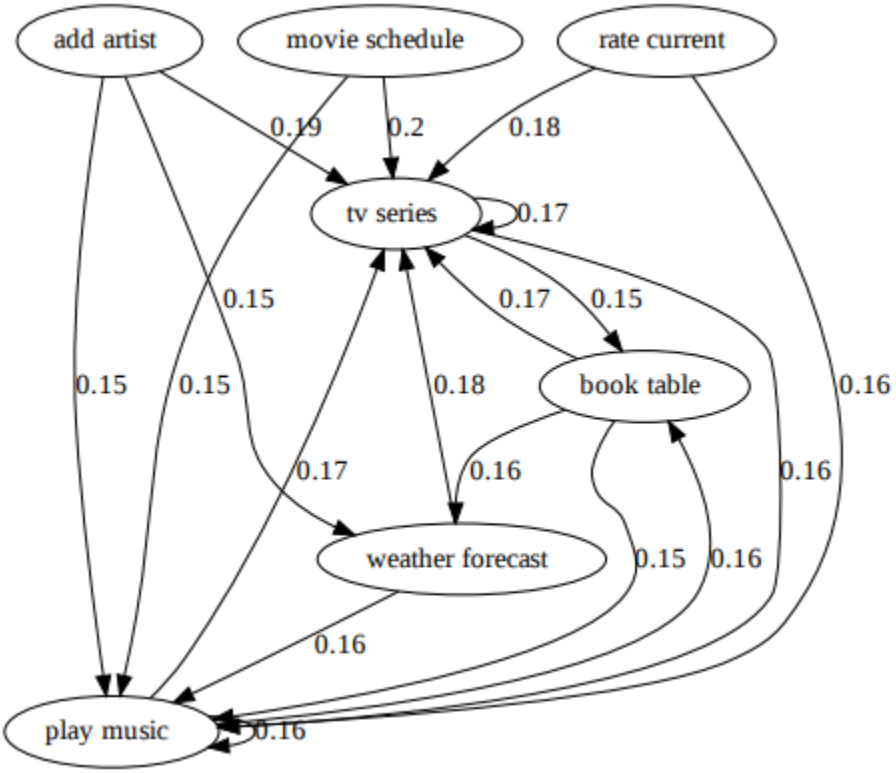


Figura 5.5 - Fluxo de diálogo SNIPS, com Bigrama

Estes primeiros fluxos não contêm o *Start of Dialogue* e o *End of Dialogue*, nem a divisão entre falas de utilizador, neste caso entre sistema e utilizador. Alguns corpora não apresentam um atributo que diferencie os utilizadores nas conversas, outros não apresentam o id de diálogo, não sendo possível assim diferenciar os diálogos e consequentemente criar o início e fim de diálogo.

O gráfico na Figura 5.6 representa o fluxo descoberto para MultiWOZ com as etiquetas dos predicados e o gráfico na Figura 5.7 representa o fluxo descoberto para SGD com os rótulos obtidos para bigramas, tendo em conta que os nós a vermelho são do utilizador, os nós a azul são do sistema e os nós amarelos correspondem ao início e ao fim do diálogo.

Para o corpus MultiWOZ de *intents* e utilizando a técnica de predicado como modo de representação, obteve-se o fluxo presente na Figura 5.6.

Para o corpus SGD de atos de diálogo e utilizando a técnica de Bigramas como modo de representação, obteve-se o fluxo presente na Figura 5.7.

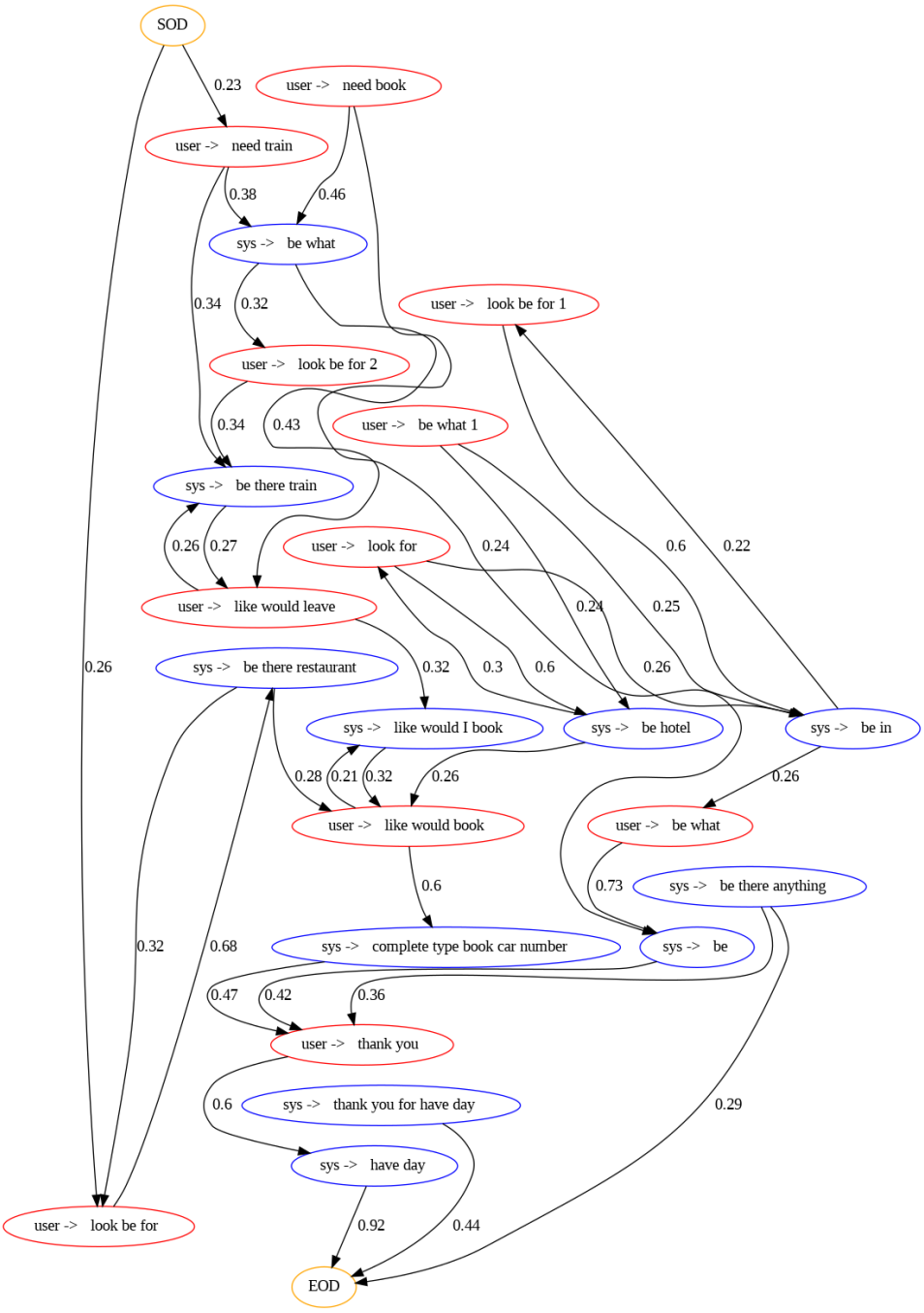


Figura 5.6 - Fluxo de diálogo MultiWOZ de *intents*, com predicado

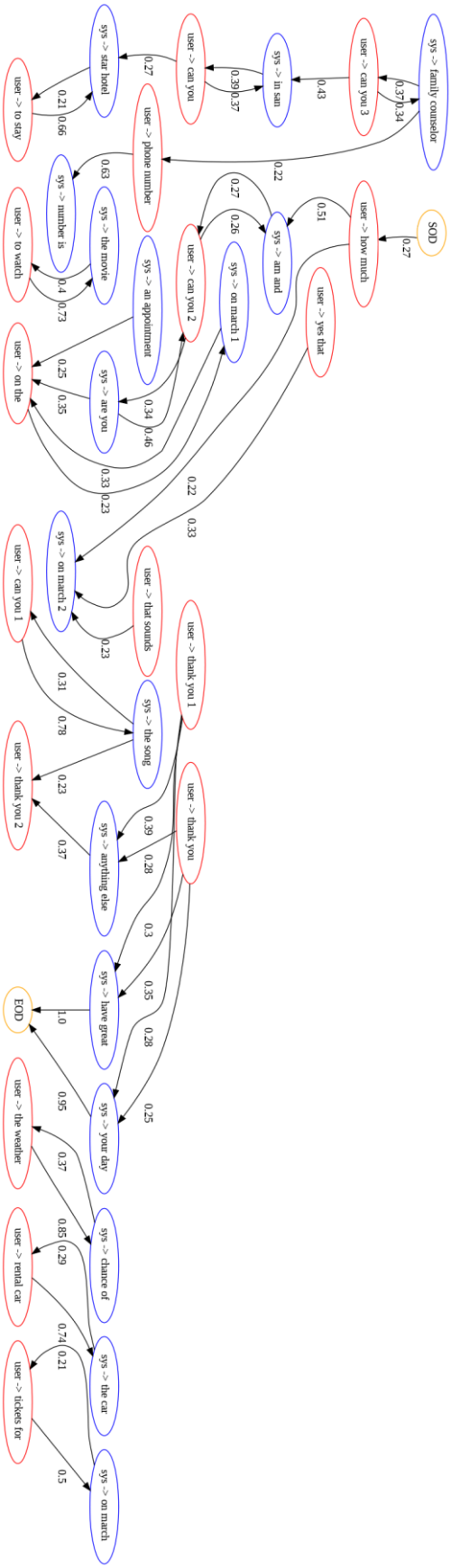


Figura 5.7 - Fluxo de diálogo SGD de atos de diálogo, com Bigramas

Estes dois últimos fluxos são muito mais conclusivos e com melhores indicações que os anteriores. Por exemplo na Figura 5.6 o nó *"like would book"* do utilizador remete para uma intenção de reservar, ou marcar algo, e tem uma probabilidade alta (0.6) do sistema de completar o tipo de reserva através do nó *"complete type book car"*. Ou assim que o utilizador agradece pela ajuda no sistema, (nó *"thank you"*) o sistema deseja-lhe um bom dia com uma probabilidade alta de 60% (nó *"have day"*), o que faz com que um diálogo termine, também indicado pela probabilidade de 92% de terminar o diálogo.

Na Figura 5.7 os assuntos são muito variados, o que não ajuda, mas ainda assim, há indicações interessantes, nomeadamente os agradecimentos do utilizador (nós *"thank you"* e *"thank you 1"*), indicando que o sistema resolveu o seu problema, remete para que o diálogo irá terminar com os nós *"have great"* e *"your day"* desejando assim um bom dia ao utilizador, terminando o diálogo a partir desses nós com 100% e 95% de probabilidades respetivamente.

5.2 Chatbot

Após a etiquetagem e a obtenção dos respetivos fluxos, foi desenvolvido um agente inteligente capaz de utilizar os fluxos como orientação relativamente a uma conversa com um humano.

Inicialmente foram feitos alguns testes com os *clusters* produzidos, mas não se estava a obter coerência nas conversas, e um dos aspetos considerados para isto acontecer foi o facto de não se separar as conversas entre os diferentes utilizadores, sejam conversas entre dois utilizadores, ou um utilizador e o sistema. Esta separação só pode ser feita em corpora que contenham falas pertencentes a mais do que um utilizador e que existe essa *feature* que distinga isso mesmo, algo que acontecia em alguns dos corpora utilizados. Por exemplo, numa conversa entre o utilizador 1 e o sistema, serão apresentados o dobro dos *clusters*, e um *cluster* do utilizador 1 iria estar ligado ao *cluster* do sistema. No caso em que os *clusters* têm falas dos dois utilizadores, é provável acontecer a resposta mais provável a uma fala do utilizador 1 ser uma fala do mesmo utilizador, o que não é coerente, e para combater isso dividiu-se os *clusters* entre os utilizadores.

Foi utilizada a estratégia para acrescentar mais dois *clusters*, um para o início de diálogo e outro a remeter para o fim de diálogo (Ritter et al., 2010). Isto mais uma vez necessita de atributos para a sua criação, já existentes em alguns corpora, como o id de turno e o id de diálogo. O id de turno, é um número incremental a cada fala pertencente ao mesmo diálogo, e o id de diálogo é um número constante no mesmo diálogo e que vai incrementando numa mudança de diálogo.

A estratégia para criar o *cluster* de início de diálogo foi de percorrer todos os diálogos e simplesmente verificar quando o id de turno era 0 (ou seja, começava a conversa)

e contar todas os *clusters* em que o id de turno era 0, conseguindo assim saber-se o início de diálogo. Para o fim de diálogo a ideia de pensamento é idêntica, encontramos o turno 0 (início de diálogo) e vamos ao anterior, o que será sempre o último turno de diálogo, uma vez que há diálogos que podem ter apenas dois turnos como ter dezasseis turnos, e mais uma vez, fazendo a contagem desses acontecimentos descobre-se as probabilidades de cada transição para o término do diálogo.

Estando criado e calculado os *clusters* para início e fim de diálogo, estes são inseridos numa matriz de ocorrências. Através dessa matriz calcula-se uma outra matriz de probabilidades, em que cada linha tem de ter uma probabilidade igual a um. Assim através dessa matriz é possível calcular o fluxo.

Tendo o fluxo de *clusters* desenvolvido e bem automatizado, segue-se a criação do *chatbot* propriamente dito. Inicia com o pedido para inserir um *input* por parte do utilizador, que é transformado em formato simples de fala, para que uma das técnicas de *Word Embedding* o receba, codifique e o transforme num vetor. Este vetor de números que corresponde ao *input* é introduzido no modelo de ML que foi utilizado anteriormente para processar e criar os *clusters* do utilizador, e dessa forma ser-lhe-á atribuído um dos *clusters* representados no fluxo. Estando já predito a que *cluster* pertence, é necessário prever qual a resposta correspondente. Existe uma estrutura de dados (*DataFrame* em Python) criada anteriormente, com a coluna com o número do *cluster*, a sua etiqueta, as falas pertencentes a esse *cluster* e qual o *cluster* mais provável de acordo com a Cadeia de Markov. Sabendo qual o *cluster* que a fala de *input* pertence, acede-se a este *DataFrame* e vê-se qual o *cluster* mais provável de se ligar. As falas que pertencem a esse *cluster* estarão codificadas com a mesma técnica de *Word Embedding* utilizada anteriormente, e irá ser calculado o cosseno de similaridade[25] entre o *embedding* de *input* e o *embedding* de todas as falas pertencentes ao *cluster* mais provável. O valor mais alto obtido de similaridade é a resposta que o *chatbot* irá dar ao utilizador, passando a ser pedido um novo *input* ao utilizador.

O *chatbot* foi testado com base no fluxo da Figura 5.8.

[25]Calcula o cosseno de similaridade entre dois exemplos. Pode encontrar-se aqui: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html [março 2023]

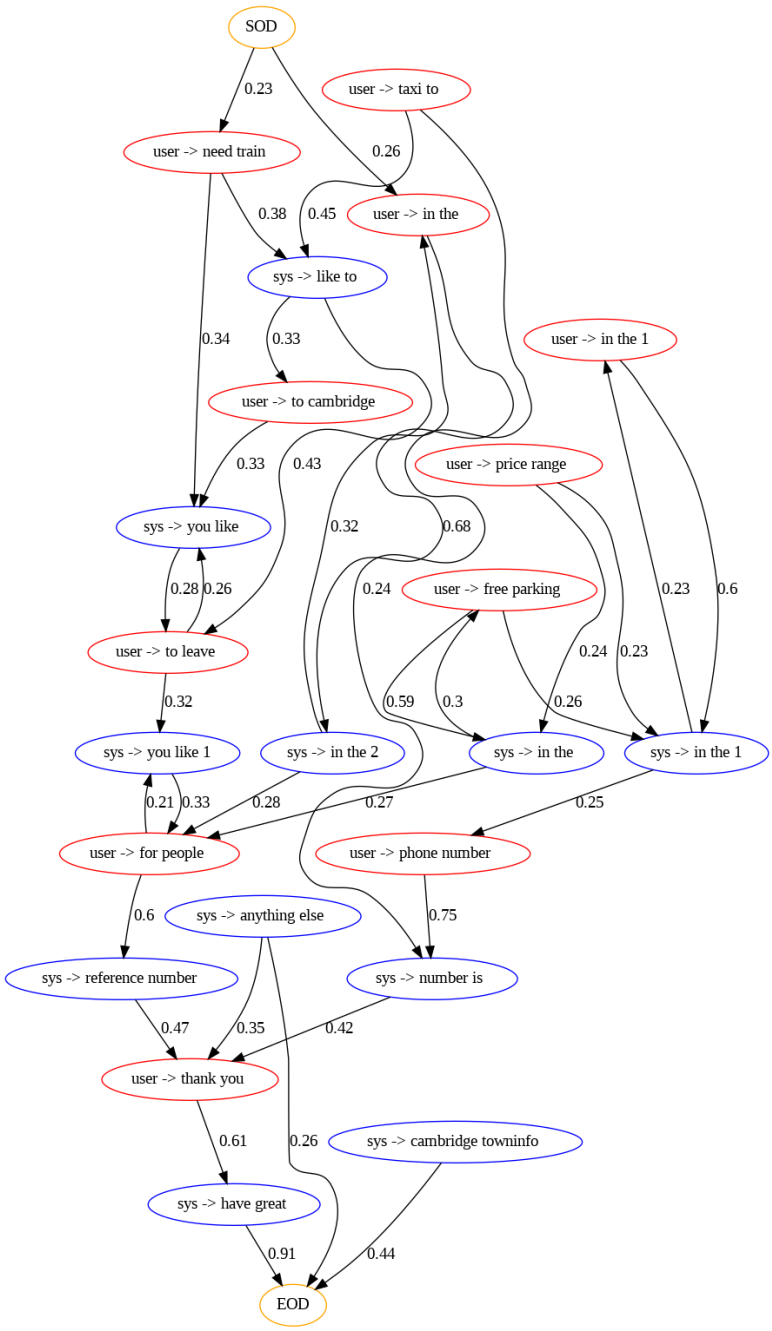


Figura 5.8 - Fluxo obtido do MultiWOZ com Bigramas para teste de *chatbot*

O utilizador começa por pedir um comboio.

User: Hello. I need a train tomorrow to London.

(*Cluster* predito foi o (*user -> need train*). Logo a resposta do *chatbot* será a fala mais próxima pertencente ao *cluster sys -> like to* pois é o *cluster* mais provável com 38%.)

Chatbot: How about the TR5015 train to Cambridge from London Liverpool? It arrives by 15:07.

Neste caso é sugerido um comboio no sentido contrário do que se pretende. Mas é dada uma resposta sobre informações de comboios e sobre Londres.

O utilizador pede então um número de telemóvel.

User: Can I get the phone number.

(*Cluster* predito foi o (*user -> phone number*). Logo a resposta do *chatbot* será a fala mais próxima pertencente ao *cluster sys -> number is* pois é o *cluster* mais provável com 75%.)

Chatbot: The area is centre and the phone number is 01223334900. Is there anything else I can help you with?

É pedido um número de telemóvel e o *chatbot* fornece, mas é dado um número sem qualquer nexo. Esta é uma abordagem simples, algo que se poderá utilizar para criar robustez seria o uso de um corpus mais genérico, por exemplo o (Nekvinda & Dušek, 2021). Visto assim sem valores específicos, poder dar respostas de acordo com o pedido pelo utilizador.

Um agradecimento.

User: Thank you so much!

(*Cluster* predito foi o (*user -> thank you*). Logo a resposta do *chatbot* será a fala mais próxima pertencente ao *cluster sys -> have great p*ois é o *cluster* mais provável com 61%.)

Chatbot: Please call us back if we can assist you in any other way. Have a great day! Goodbye.

E de seguida terminaria a conversa, pois partindo do *cluster sys -> have great* tem-se 91% de probabilidade de ir para o fim do diálogo.

Aqui foi o exemplo de uma resposta perfeita por parte do *chatbot*.

6 CONCLUSÕES E TRABALHO FUTURO

Foi adotada uma abordagem não supervisionada para comparar anotações de atos de diálogo e de *intents*, de vários conjuntos de dados que contêm falas com base em diferentes representações. A plataforma desenvolvida descobre e extrai fluxos de diálogo a partir de um histórico, independentemente do domínio ou de anotações. Esses fluxos conseguem ser facilmente interpretáveis pelos humanos ou agentes inteligentes.

As falas foram representadas de várias formas, tendo os modelos de *Sentence Transformer* obtido os melhores resultados globalmente, pois aproximaram-se mais dos atos de diálogo e dos *intents* para os diferentes corpora.

As experiências realizadas mostraram que os resultados obtidos através de *clustering* não são diretamente atos de diálogo ou *intents*, porém, através da etiquetagem dos *clusters*, foi possível verificar que uns poderão ser compostos por atos de diálogo, enquanto outros representam *intents*. Exemplo disso são os *clusters "thank you"* e *"need train"* da Figura 5.8 que traduzem um ato de diálogo e um *intent*, respetivamente.

Os fluxos resultantes fornecem uma visão geral de alto nível dos diálogos, que podem ter diferentes aplicações, por exemplo em serviços de segurança social de forma a ajudar os cidadãos a encontrar as informações de maneira mais eficiente e facilitar o atendimento ao cliente, ou sistemas de saúde 24 horas, ara fornecer suporte médico a pacientes ou ajudar na sua triagem.

O trabalho desenvolvido poderá ser aprofundando através de abordagens distintas, como a utilização de diferentes formas de representação de falas, ou outros métodos para etiquetar os *clusters*. O *chatbot* desenvolvido neste projeto, apesar de simples, ajuda a mostrar como um agente, artificial ou humano, pode tirar partido dos fluxos descobertos. No entanto, poderá ser aprimorado de forma a criar um sistema com mais robustez e até mesmo ter a capacidade de dar respostas através do histórico de diálogos e conexões entre palavras para formar uma resposta coesa e com nexo, dando respostas inseridas no mesmo domínio. Para tal pode ser utilizado um corpus mais genérico.

Num balanço final, após meses a investigar um paradigma, posso afirmar que este projeto me trouxe desafios a nível pessoal e profissional capazes de me tornar mais qualificado e habilitado para futuros reptos. Além disso, ao longo do tempo, descobri uma área que é realmente do meu interesse e que terei curiosidade para continuar a aprender e desenvolver.

REFERÊNCIAS BIBLIOGRÁFICAS

Austin, J. L. (1975). *How To Do Things With Words: The William James Lectures delivered at Harvard University in 1955*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198245537.001.0001

Ayanouz, S., Abdelhakim, B. A., & Benhmed, M. (2020, March 31). A Smart Chatbot Architecture based NLP and Machine Learning for Health Care Assistance. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3386723.3387897

Boné, J., Ferreira, J. C., Ribeiro, R., & Cadete, G. (2020). Disbot: A Portuguese disaster support dynamic knowledge chatbot. *Applied Sciences (Switzerland)*, *10*(24), 1–20. https://doi.org/10.3390/app10249082

Bouraoui, J.-L., & Lemaire, V. (2017). *Cluster-Based Graphs for Conceiving Dialog Systems*.

Bouraoui, J. L., Le Meitour, S., Carbou, R., Rojas-Barahona, L. M., & Lemaire, V. (2019). Graph2Bots, unsupervised assistance for designing chatbots. *SIGDIAL 2019 - 20th Annual Meeting of the Special Interest Group Discourse Dialogue - Proceedings of the Conference*, 114–117. https://doi.org/10.18653/v1/w19-5915

Bradeško, L., & Mladenić, D. (2012). A Survey of Chatbot Systems through a Loebner Prize Competition A Survey of Chabot Systems through a Loebner Prize Competition. *Proceedings of Slovenian Language Technologies Society Eighth Conference of Language Technologies*, *2*(October 2012), 34–37. http://www-ai.ijs.si/eliza-cgi-bin/eliza_script

Budzianowski, P., Wen, T. H., Tseng, B. H., Casanueva, I., Ultes, S., Ramadan, O., & Gašić, M. (2018). MultiWoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, 5016–5026. https://doi.org/10.18653/v1/d18-1547

Bunt, H., Petukhova, V., Malchanau, A., Fang, A., & Wijnhoven, K. (2016). The DialogBank. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 3151–3158.

Casanueva, I., Temčinas, T., Gerz, D., Henderson, M., & Vulić, I. (2020). *Efficient Intent Detection with Dual Sentence Encoders*. 38–45. https://doi.org/10.18653/v1/2020.nlp4convai-1.5

Cerisara, C., Jafaritazehjani, S., Oluokun, A., & Le, H. T. (2018). Multi-task dialog act and sentiment recognition on Mastodon. *COLING 2018 - 27th International*

Conference on Computational Linguistics, Proceedings, 745–754.

Chen, C.-Y., Yu, D., Wen, W., Yang, Y. M., Zhang, J., Zhou, M., Jesse, K., Chau, A., Bhowmick, A., Iyer, S., Sreenivasulu, G., Cheng, R., Bhandare, A., & Yu, Z. (2018). Gunrock: Building A Human-Like Social Bot By Leveraging Large Scale Real User Data. *Proceedings of the 2018 Amazon Alexa Prize*. http://aws.amazon.com/lambda

Collins, M. J. (1996). *A new statistical parser based on bigram lexical dependencies*. 184–191. https://doi.org/10.3115/981863.981888

Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D., Doumouro, C., Gisselbrecht, T., Caltagirone, F., Lavril, T., Primet, M., & Dureau, J. (2018). *Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces*. https://doi.org/10.48550/arxiv.1805.10190

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Duran, N., Battle, S., Smith, J., & Duran, N. (2021). Sentence encoding for Dialogue Act classification. *Natural Language Engineering*. https://doi.org/10.1017/S1351324921000310

Eisenstein, J. (2018). *Eisenstein NLP Notes*.

Ezen-Can, A., & Boyer, K. E. (2013). Unsupervised classification of student dialogue acts with query-likelihood clustering. *Proceedings of the 6th International Conference on Educational Data Mining, EDM 2013*.

Ezen-Can, A., & Boyer, K. E. (2015). Understanding Student Language: An Unsupervised Dialogue Act Classification Approach. *EDM 2015*.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *1*, 517–520. https://doi.org/10.1109/ICASSP.1992.225858

Gonçalo Oliveira, H., Ferreira, P., Martins, D., Silva, C., & Alves, A. (2022). A Brief Survey of Textual Dialogue Corpora. *Proceedings of the Language Resources and Evaluation Conference*, *676*(June), 1264–1274. https://aclanthology.org/2022.lrec-1.135

Griol, D., Callejas, Z., Molina, J. M., & Sanchis, A. (2021). Adaptive dialogue management using intent clustering and fuzzy rules. *Expert Systems*, *38*(1). https://doi.org/10.1111/exsy.12630

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure,

dynamics, and function using NetworkX. In G. Varoquaux & Travis Vaught (Eds.), *7th Python in Science Conference (SciPy 2008)* (pp. 11–15).

Jollife, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (Vol. 374, Issue 2065). Royal Society of London. https://doi.org/10.1098/rsta.2015.0202

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. In *Journal of Documentation* (Vol. 28, Issue 1, pp. 11–21). https://doi.org/10.1108/eb026526

Joty, S., Carenini, G., & Lin, C. Y. (2011). Unsupervised modeling of dialog acts in asynchronous conversations. *IJCAI International Joint Conference on Artificial Intelligence*. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-303

Li, X., Ouyang, J., Zhou, X., Lu, Y., & Liu, Y. (2015). Supervised labeled latent dirichlet allocation for document categorization. *Applied Intelligence*, *42*(3), 581–593. https://doi.org/10.1007/s10489-014-0595-0

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). *DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset*. 986–995. http://arxiv.org/abs/1710.03957

Liu, P., Ning, Y., Wu, K. K., Li, K., & Meng, H. (2021). *Open Intent Discovery through Unsupervised Semantic Clustering and Dependency Parsing*. arXiv. https://doi.org/10.48550/ARXIV.2104.12114

Liu, Y., Han, K., Tan, Z., & Lei, Y. (2017). Using context information for dialog act classification in DNN framework. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2170–2178. https://doi.org/10.18653/v1/d17-1231

Luhn, H. P. (2010). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, *1*(4), 309–317. https://doi.org/10.1147/rd.14.0309

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 281–296.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations ofwords and phrases and their compositionality. *Advances in Neural Information Processing Systems*.

Mou, Y., He, K., Wu, Y., Zeng, Z., Xu, H., Jiang, H., Wu, W., & Xu, W. (2022). Disentangled Knowledge Transfer for OOD Intent Discovery with Unified Contrastive Learning. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 46–53.

https://doi.org/10.18653/v1/2022.acl-short.6

Nekvinda, T., & Dušek, O. (2021). Shades of BLEU, Flavours of Success: The Case of MultiWOZ. *GEM 2021 - 1st Workshop on Natural Language Generation, Evaluation, and Metrics, Proceedings*, 34–46. https://doi.org/10.18653/v1/2021.gem-1.4

Park, J., Jang, Y., Lee, C., & Lim, H. (2022). *Analysis of Utterance Embeddings and Clustering Methods Related to Intent Induction for Task-Oriented Dialogue*. https://doi.org/10.48550/arxiv.2212.02021

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1532–1543. https://doi.org/10.3115/v1/d14-1162

Penta, A., & Pal, A. (2021). What is this Cluster about? Explaining textual clusters by extracting relevant keywords. *Knowledge-Based Systems*, *229*. https://doi.org/10.1016/j.knosys.2021.107342

Popescu-Belis, A. (2005). Dialogue acts: One or more dimensions. *ISSCO WorkingPaper62*, *DECEMBER 2005*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.127.5668&rep=rep1&type=pdf

Rakesh Kumar Sharma. (2020). An Analytical Study and Review of open source Chatbot framework, Rasa. *International Journal of Engineering Research And*, *V9*(06). https://doi.org/10.17577/IJERTV9IS060723

Rastogi, A., Zang, X., Sunkara, S., Gupta, R., & Khaitan, P. (2020). Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 8689–8696. https://doi.org/10.1609/aaai.v34i05.6394

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 3982–3992. https://doi.org/10.18653/v1/d19-1410

Ribeiro, E., Ribeiro, R., & De Matos, D. M. (2019). Deep dialog act recognition using multiple token, segment, and context information representations. *Journal of Artificial Intelligence Research*, *66*, 861–899. https://doi.org/10.1613/jair.1.11594

Ribeiro, E., Ribeiro, R., & Matos, D. M. de. (2019). Reconhecimento de Actos de Diálogo Hierárquicos e Multi-Etiqueta em Dados em Espanhol. *Linguamática*,

11(1), 17–40. https://doi.org/10.21814/lm.11.1.278

Ritter, A., Cherry, C., & Dolan, B. (2010). Unsupervised modeling of twitter conversations. *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, 172–180.

Rosenberg, A., & Hirschberg, J. (2007). V-Measure: A conditional entropy-based external cluster evaluation measure. *EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 410–420.

Rousseeuw, P. J. (1987a). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Rousseeuw, P. J. (1987b). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sabharwal, N., & Agrawal, A. (2020). Cognitive Virtual Assistants Using Google Dialogflow. In *Cognitive Virtual Assistants Using Google Dialogflow*. Apress. https://doi.org/10.1007/978-1-4842-5741-8

Serradilla, O., Zugasti, E., Rodriguez, J., & Zurutuza, U. (2022). Deep learning models for predictive maintenance: a survey, comparison, challenges and prospects. *Applied Intelligence*, *52*(10), 10934–10964. https://doi.org/10.1007/s10489-021-03004-y

Shang, L., Lu, Z., & Li, H. (2015). Neural responding machine for short-Text conversation. *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference*, *1*, 1577–1586. https://doi.org/10.3115/v1/p15-1152

Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Van Ess-Dykema, C., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, *26*(3), 339–373. https://doi.org/10.1162/089120100561737

Truong, K. N., Huang, E. M., & Abowd, G. D. (2004). CAMP: A magnetic poetry interface for end-user programming of capture applications for the home. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3205*, 143–160. https://doi.org/10.1007/978-3-540-30119-6_9

Van Der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2625.

Vedula, N., Lipka, N., Maneriker, P., & Parthasarathy, S. (2020). Open Intent Extraction from Natural Language Interactions. *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020*, 2009–2020. https://doi.org/10.1145/3366423.3380268

Vysala, A., & Gomes, D. J. (2020). *Evaluating and Validating Cluster Results*. 37–47. https://doi.org/10.5121/csit.2020.100904

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2021). "Let me explain!": exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces*, *15*(2), 87–98. https://doi.org/10.1007/s12193-020-00332-0

Weizenbaum, J. (1966). ELIZA-A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Wen, T. H., Vandyke, D., Mrkšíc, N., Gašíc, M., Rojas-Barahona, L. M., Su, P. H., Ultes, S., & Young, S. (2017). A network-based end-to-end trainable task-oriented dialogue system. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, *1*, 438–449. https://doi.org/10.18653/v1/e17-1042

Yan, Z., Duan, N., Chen, P., Zhou, M., Zhou, J., & Li, Z. (2017). Building task-oriented dialogue systems for online shopping. *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 4618–4625. https://doi.org/10.1609/aaai.v31i1.11182

Zhang, H., Xu, H., Lin, T. E., & Lyu, R. (2021). Discovering New Intents with Deep Aligned Clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(16), 14365–14373. https://doi.org/10.1609/AAAI.V35I16.17689

ANEXO A - PROPOSTA DE ESTÁGIO

DEPARTAMENTO DE ENGENHARIA
INFORMÁTICA E DE SISTEMAS

PROPOSTA DE ESTÁGIO

Ano Letivo de 2021/2022

Mestrado em Engenharia Informática

(Engenharia de Software ou Análise Inteligente de Dados)

TEMA

Classificação Automática de Fluxos de Diálogo (@FLOWANCE)

SUMÁRIO

Fruto da transformação digital e da emergência de novas tecnologias, os centros de contacto têm sofrido uma enorme pressão nos últimos anos. A expetativa dos clientes quando interagem com um centro de contacto no que respeita o tempo de resposta e a capacidade de resolução dos seus problemas é cada vez mais exigente. Sendo os agentes (operadores humanos) centrais na excelência da experiência de contacto de clientes com o centro de atendimento, a indústria tem investido fortemente em processos de assistência em tempo-real com recurso a Inteligência Artificial (IA) e Machine Learning (ML).
Este trabalho tem como objetivo explorar abordagens para classificar e agrupar segmentos de diálogos, e assim chegar a representações que possam ser úteis a sistemas de orientação do assistente humano. Ainda que o principal objetivo deste trabalho seja desenvolver técnicas para apoio a assistentes humanos, muitas delas poderão também contribuir para o desenvolvimento de agentes artificiais (chatbots) cada vez mais inteligentes.

1. ÂMBITO

É cada vez mais comum recorrer a técnicas de Inteligência Artificial e, especificamente, Processamento de Linguagem Natural (PLN) no atendimento remoto de clientes. No caso dos chatbots, o atendimento pode nem precisar da intervenção de um agente humano. Mas também é possível tirar partido de ferramentas inteligentes

para assistir agentes humanos, com vista a tornar as conversas mais eficientes, sem perder a naturalidade.

Entre essas ferramentas incluem-se sistemas de orientação ou de recomendação. Com base na monitorização da conversa e em conhecimento aprendido com conversas anteriores, estes sistemas podem prever interações futuras e guiar o assistente humano, tanto na obtenção de melhores respostas, como na formulação ou sugestão de perguntas a realizar, tendo também como objetivo maximizar a satisfação do cliente.

2. OBJETIVOS

Este trabalho tem como objetivo principal explorar abordagens para classificar e agrupar segmentos de diálogos, e assim chegar a representações que possam ser úteis a sistemas de orientação.

Isto deverá envolver um levantamento e exploração de abordagens para objetivos intermédios tais como:

- Descoberta de segmentos ou atos comuns a diferentes diálogos (e.g., pergunta, resposta, opinião);

- Classificação automática de atos de diálogo de acordo com uma lista pré-definida.

- Identificação de sequências de atos comuns a um conjunto de diálogos (fluxos).

O ponto de partida serão conjuntos de diálogos disponíveis de forma aberta para investigação científica.

Em qualquer um dos casos, deverá ser considerado um pré-processamento de dados, com recurso a ferramentas de PLN, e que ajude na abstração de caraterísticas associadas ao assunto de caraterísticas efetivamente úteis à classificação dos atos de diálogo.

Finalmente, os fluxos deverão emergir da análise de padrões comuns num conjunto de diálogos já anotados.

3. PROGRAMA DE TRABALHOS

O estágio consistirá nas seguintes atividades e respetivas tarefas:

- ***T1*** *- Revisão da Literatura* - estudo de conceitos e trabalhos relacionados em PLN, Análise e Modelação de Diálogo;
- ***T2*** *- Identificação de dados disponíveis e familiarização com ferramentas a utilizar;*

- ***T3*** *- Experiências iniciais (e.g., réplica de um trabalho relacionado que classifique atos de diálogo num novo conjunto de dados);*
- ***T4*** *- Classificação automática de atos de diálogo;*
- ***T5*** *- Agrupamento de atos diálogo em conjuntos de dados não-anotados;*
- ***T6*** *- Descoberta de fluxos de diálogo;*
- ***T7*** *- Avaliação de soluções*
- ***T8*** *- Elaboração da dissertação de mestrado*
- ***T9*** *- Escrita de artigo científico.*

4. CALENDARIZAÇÃO DAS TAREFAS

As Tarefas acima descritas, incluindo os testes de validação de cada módulo, serão executadas de acordo com a seguinte calendarização:

O plano de escalonamento dos trabalhos é apresentado em seguida:

A	B	C	D	E	F	G	H	I	J	K
	Meses									
Tarefas	1	2	3	4	5	6	7	8	9	10
T1										
T2										
T3										
T4										
T5										
T6										
T7										
T8										
T9										
Metas	INI		M1	M2		M3	M4	M5	M6	M7

INI Início dos trabalhos
M1 Tarefas T1 e T2 terminadas
M2 Tarefa T3 terminada
M3 Tarefa T4 terminada
M4 Tarefa T5 terminada
M5 Tarefa T6 terminada
M6 Tarefa T7 terminada
M7 Tarefas T8 e T9 terminadas

DEPARTAMENTO DE ENGENHARIA
INFORMÁTICA E DE SISTEMAS

5. RESULTADOS

Os resultados do estágio serão consubstanciados num conjunto de documentos, a elaborar pelo estagiário de acordo com o seguinte plano:

- Descrição do estado da arte
- Proposta do método de classificação de atos de diálogo
- Proposta do método de agrupamento de atos de diálogo
- Análise comparativa de cada método
- Dissertação de mestrado
- Artigo científico

6. LOCAL DE TRABALHO

O local de trabalho será em regime híbrido (remoto/presencial num laboratório do CISUC, no DEI, onde haverá um acompanhamento regular por parte dos orientadores).

7. METODOLOGIA

Reuniões semanais e elaboração de um dossier de estágio numa pasta partilhada com a bibliografia anotada, documentação produzida e código implementado.

8. ORIENTAÇÃO

DEIS/ISEC:
Ana Alves (aalves@isec.pt)
Professor Adjunto

DEI/CISUC:
Hugo Gonçalo Oliveira (hroliv@dei.uc.pt)
Professor Auxiliar

9. CARACTERIZAÇÃO E REMUNERAÇÃO

- Data de início 04/10/2021
- Data de fim 29/07/2021

DEPARTAMENTO DE ENGENHARIA
INFORMÁTICA E DE SISTEMAS

- O trabalho será integrado num projeto de investigação em co-promoção, no âmbito do qual o estudante se poderá candidatar a uma bolsa de investigação para licenciado, no segundo semestre.

ANEXO B - ARTIGO PUBLICADO - RECPAD 22

Can we rely on clustering for approximating intents and dialog acts?

Daniel Martins[12]
daniel.martins2997@gmail.com
Patrícia Ferreira[12]
patriciaf@dei.uc.pt
Hugo Gonçalo Oliveira[23]
hroliv@dei.uc.pt
Catarina Silva[23]
catarina@dei.uc.pt
Ana Alves[12]
ana@dei.uc.pt

[1] ISEC, Instituto Politécnico de Coimbra, Portugal
[2] CISUC, Universidade de Coimbra, Portugal
[3] DEI, FCTUC, Universidade de Coimbra, Portugal

Abstract

The classification of dialog acts and intents can play an important role in many natural language applications, notably in the conversation-based scenarios as chatbots. In this work the goal is to recognize and group dialog acts and/or intents using a clustering approach. The strategy is domain-independent, presenting the potential to enhance and support different application scenarios. Results show that clusters are closer to intents, presenting better results by external evaluation. Although more experiments should be carried out, the labels automatically chosen for intent clusters show clearer meaning than dialog act clusters labels, since the latter present greater entropy for different types of dialog acts.

1 Introduction

There are usually high expectations regarding customer-service call-centers, where fast responses are needed, in order to solve problems. Motivated by these requirements, we introduce some methods in attempt to automatically approximate to intents and dialog acts using unsupervised learning. This work can be the starting point to develop agents that follow paths influenced by the utterances that were grouped in clusters.

Dialog acts reveal the intention behind the words of their interlocutors; with intents, it boils down to the primary goal of the speakers and that is important to achieve the communicative intention behind each utterance. In sum, dialog acts can be stated to understand the intention of the user and intents represent the primary goal of the user. For example, the utterance: *"Can I have a pizza?"*, may have a dialog act and a intent. The user's aim is to ask a question, so the dialog act might be labelled as *request*. Regarding the intent, the speaker's main goal is to order food, so could be labelled as *order_food*.

Based on intent recognition, Casanueva et al. [2] used pre-trained Sentence Encoding models, as will be applied in this work with Sentence Transformer, they used a dual encoder model, a straightforward combination of two pretrained dual sentence encoders (USE+ConveRT) and separately. In an attempt to assess whether the combination of USE+ConveRT would bring better results in the intent recognition task, the results showed that the combination of the two models has better accuracy scores in almost all testes in the different corpora used comparing to the results of BERT, USE and ConveRT models.

Wang et al. [7] sought to understand the response of using the historical responses of the interlocutors, in an attempt to efficiently group the acts of dialogue into clusters. They concluded that using models that consider the history of conversations, by using a Heterogeneous User History graph, and the results us superior compared to other models.

The remainder of the paper is organized as follows: Section 2 describes the approach with all steps described used in this work; in Section 3 the results obtained are described and discussed; finally, Section 4 presents the conclusions and a few guidelines for future work.

2 Proposed Approach

This work tested different Natural Language Processing methods in different corpora to understand which methods are most efficient in certain corpus and why. The used corpora are all tagged for validation and further evaluation, and tests were performed for two corpora with intents and two corpora with dialog acts.

The proposed approach consists of several steps:

1. **Corpora Selection**: Choose public corpora with intent or dialog act labelled data.
2. **Pre-Processing**: Use NLP methods to prepare text data from selected corpus for the model building..
3. **Feature Engineering**: Transform text into numeric vectors using word embedding techniques
4. **Unsupervised Learning**: Perform K-Means clustering
5. **Evaluate and Validate**: Use of metrics to evaluate if clusters come close to dialog acts and / or intents.

2.1 Corpora

Within the scope of the recognition of acts of dialogue, the corpora chosen was CamRest676 [8] and Mastodon [3]. For the recognition of intents the MultiWOz [1] and Snips [4] were chosen.

CamRest676: Multi-label problem; 5488 utterances about restaurants information, labelled with 16 type of dialog acts;

Mastodon: Single-label problem; 2203 utterances taken from social networks, labelled with 8 type of dialog acts;

MultiWOZ: Multi-label problem; 2987 multi-domain utterances labelled with 11 type of intents;

Snips: Single-label problem; 13804 multi-domain utterances labelled with 7 type of intents;

2.2 Methodology

All sentences from the different corpora were subjected to some simple preprocessing techniques, like lowercase transformation or removal of stopwords. Since the text is all processed, it is necessary to transform it into numerical vectors, for which three different techniques were used:

1. TF-IDF: a statistical measure that evaluates how relevant a string representation (words, phrases, lemmas, etc.) is to a document in a collection of documents. This method comes to a conclusion by multiplying how many times a word appears in a document, and the inverse document frequency of the word across a set of documents .

2. Word2vec: a method to efficiently create word embeddings. It takes a text corpus as input and the word vectors creation process is performed by determining which words the target word occurs with more often, and show the semantic closeness between the words [5].

Word2vec produces one vector per word, whereas TF-IDF produces a score.

3. Sentence Transformer: a framework with a large collection of pre-trained models which can be easily used for text embedding. It can compute dense vector representations for sentences and the text is embedding in vector space such that similar text is close and can efficiently be found using cosine similarity [6].

After using one of these different techniques, the next step is clustering by using K-Means. For this work, the default parameters were used, while the number of clusters (k) corresponds to the number of dialog acts/intents through the corpus used.

The results obtained from clustering will be evaluated internally and externally. The internal evaluation is focused on the data that will be

clustered, the quality of the data. The metrics of this evaluation method to be used are the Silhouette Score that compares the average distance to elements in other clusters with elements in the same cluster (maximum class spread/variance) and Davies–Bouldin Index (DBI) (maximum interclass-intraclass distance ratio). The external evaluation is based on the clusters that were produced, so this evaluation requires the existence of pre-labelled data, this measures how close the clustering is to the pre-labelled data. For this evaluation method it will be applied the V-Measure metric that is the harmonic mean between homogeneity and completeness and Accuracy that try to find the best match between the class labels and the cluster labels.

3 Results and Discussion

As already mentioned, the three sentence/word embedding techniques were tested in the four corpora (two with intents and two with dialog acts), taking into account the four evaluation metrics. The results are presented in Tables 1, 2, 3 and 4:

Table 1: DAs Clustering - Applying techniques on CamRest676

Representation Model	Evaluation			
	Internal		External	
	Silhouette Score	DBI	Accuracy	V-Measure
TF-IDF	**0.167**	3.066	0.049	**0.134**
Sentence Transformer	0.104	**2.624**	**0.051**	0.123
word2vec	0.098	3.732	0.046	0.121

Table 2: DAs Clustering - Applying techniques on Mastodon

Representation Model	Evaluation			
	Internal		External	
	Silhouette Score	DBI	Accuracy	V-Measure
TF-IDF	**0.099**	**2.661**	0.039	0.042
Sentence Transformer	0.018	5.433	**0.133**	**0.069**
word2vec	0.087	3.432	0.0264	0.037

Table 3: Intents Clustering - Applying techniques on MultiWOZ

Representation Model	Evaluation			
	Internal		External	
	Silhouette Score	DBI	Accuracy	V-Measure
TF-IDF	0.047	4.122	0.138	0.301
Sentence Transformer	**0.094**	**2.872**	**0.195**	**0.449**
word2vec	0.032	4.643	0.102	0.248

Table 4: Intents Clustering - Applying techniques on Snips

Representation Model	Evaluation			
	Internal		External	
	Silhouette Score	DBI	Accuracy	V-Measure
TF-IDF	0.044	4.776	0.21	0.64
Sentence Transformer	**0.085**	**3.605**	**0.274**	**0.814**
word2vec	0.041	4.832	0.184	0.598

In a brief analysis of the tables, it is possible to verify that the external evaluation was superior in the corpora with intents. On the other hand, the results of the internal evaluation are quite similar in the corpora of dialog acts and of intents.

The t-SNE is an excellent tool to understand high-dimensional datasets, therefore, it was applied in CamRest, as can be seen in Figure 1, and in Figure 2 for the Snips corpus.

Figure 1: t-SNE for CamRest

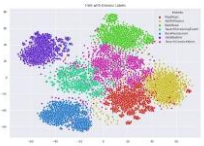


Figure 2: t-SNE for Snips

At first sight, the difference between the two figures is evident, in which the 7 Snips clusters are all very well delimited, with few outliers. Regarding CamRest, there are some classes that group together well, such as *request_postcode*, *none* and *request_food*, but all other classes are very dispersed. This graphical visualization tool was made for all corpora, and is in line with the results that were presented in tables 1 through 4, mainly by using the V-Measure metric differentiating between the intents and the dialog acts, leading to better results for the intents.

4 Conclusions and Future Work

The main general conclusions obtained by evaluating the results is that the Sentence Transformer presents best performance, and it is the most regular method throughout the different corpora.

Through analyses of the results between intents and dialog acts, there are some differences and similarities between them, but our methods are closer to discovering intents than dialog acts.

In the internal evaluation, there are no outstanding results, and they end up being similar, with a slight increase for dialog acts, but nothing conclusive. In the external evaluation, there are greater differences, with the intents getting much better results, especially in the V-Measure metric. These results are visually explained through the two t-SNE graphs presented, indicating that clustering is closer to intents, since the external evaluation takes into account the labels that are produced, and not with data, which is verified in the intents graphs.

As future work other methods should be explored, since the results obtained were not ideal. Future lines of research may focus on sentence Embedding, evaluation metrics, and algorithms. As a follow-up to this work, labelling the clusters produced can be an asset in order to see what they are composed of and what they are most similar to, whether they are intents or dialog acts.

Acknowledgements This work was financially supported by the project FLOWANCE (POCI-01-0247-FEDER-047022), cofinanced by the European Regional Development Fund (FEDER), through Portugal 2020 (PT2020), and by the Competitiveness and Internationalization Operational Programme (COMPETE 2020); and by national funds through FCT, within the scope of the project CISUC (UID/CEC/00326/2020) and by European Social Fund, through the Regional Operational Program Centro 2020.

References

[1] Paweł Budzianowski, Tsung Hsien Wen, Bo Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*. doi: 10.18653/v1/d18-1547.

[2] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient Intent Detection with Dual Sentence Encoders. pages 38–45. Association for Computational Linguistics (ACL), jul 2020. doi: 10.18653/v1/2020.nlp4convai-1.5.

[3] Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. Multi-task dialog act and sentiment recognition on Mastodon. In *COLING 2018 - 27th International Conference on Computational Linguistics, Proceedings*, pages 745–754. Association for Computational Linguistics (ACL), 2018. ISBN 9781948087506.

[4] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *CoRR*, abs/1805.10190, 2018.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. URL `https://arxiv.org/abs/1301.3781`.

[6] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL `https://arxiv.org/abs/1908.10084`.

[7] Dong Wang, Ziran Li, Haitao Zheng, and Ying Shen. Integrating User History into Heterogeneous Graph for Dialogue Act Recognition. pages 4211–4221. Association for Computational Linguistics (ACL), jan 2021. doi: 10.18653/v1/2020.coling-main.372.

[8] Tsung Hsien Wen, David Vandyke, Nikola Mrkšíc, Milica Gašíc, Lina M. Rojas-Barahona, Pei Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proc. of Conference*, volume 1, pages 438–449. ACL, 2017. doi: 10.18653/v1/e17-1042.

ANEXO C - COAUTORIA ARTIGO PUBLICADO - LREC 22

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1264–1274
Marseille, 20-25 June 2022

A Brief Survey of Textual Dialogue Corpora

Hugo Gonçalo Oliveira[1,2], **Patrícia Ferreira**[1,3], **Daniel Martins**[3], **Catarina Silva**[1,2], **Ana Alves**[1,3]
[1]CISUC, Universidade de Coimbra, Portugal
[2]DEI, FCTUC, Universidade de Coimbra, Portugal
[3]ISEC, Instituto Politécnico de Coimbra, Portugal
hroliv@dei.uc.pt, patriciaf@dei.uc.pt, daniel.martins2997@gmail.com, catarina@dei.uc.pt, ana@dei.uc.pt

Abstract

Several dialogue corpora are currently available for research purposes, but they still fall short for the growing interest in the development of dialogue systems with their own specific requirements. In order to help those requiring such a corpus, this paper surveys a range of available options, in terms of aspects like speakers, size, languages, collection, annotations, and domains. Some trends are identified and possible approaches for the creation of new corpora are also discussed.

Keywords: Dialogue Corpora, Dialogue Analysis, Dialogue Systems

1. Introduction

Driven by the growing adoption of dialogue systems and conversational agents in recent years (Dale, 2016), the need for conversational data to study specific dialogue phenomena and train the aforementioned systems has also increased. However, despite the broad range of dialogue corpora currently available, many released in the last years, this number is still far from covering all the necessities of real-world applications of dialogue systems. These include all possibly envisioned dialogue analysis tasks, domains, and their specificity, among others, such as languages. Additionally, the access to dialogue data and its usage are typically very restricted, due to the logistics involved in their collection and, especially, privacy concerns. For instance, a common application of dialogue systems is automated customer-support, e.g., through call-centers. However, real conversations of this type typically include a large amount of personal data, not always easy to anonymise, as well as sensitive information such as unsatisfied customers expressing their opinion.

Hence, in most cases, there will not be a perfect corpus for a new project involving dialogue. This means that researchers will often have to choose between: (i) going through all the work involved in the collection (and annotation) of data for creating a new corpus from scratch; or (ii) looking at publicly available corpora for selecting the most suitable for their task, possibly after some adaptations.

This survey targets both previous scenarios. It is not as extensive as Serban et al. (2018), but it gives a more practical perspective, and is more up-to-date, with a focus on corpora of human-human dialogue, available in a textual format, not exclusively in English. More precisely, we compile a list of corpora of such dialogues, and put forward a brief analysis, covering the following aspects: number and type of speakers, size, languages, annotations, data collection, and contents in general. Interested researchers may use this survey as a reference for selecting appropriate corpora for their projects, but also for deciding on suitable approaches for creating new dialogue corpora.

With this analysis, we can identify trends, which end up contributing to identify gaps. There are dialogue corpora of variable sizes, even though we could think of many more, on varied domains. The largest corpora have no annotations, whereas common annotations include dialogue acts and states, and, in some cases, also sentiment-related information. Although we target corpora available as text, some are originally spoken corpora that have been transcribed. Out of those, we focus on corpora that include linguistic annotations, not dependent on the audio. Due to the constraints of using real conversations, most of the available corpora were created in a crowdsourcing environment, where workers knew they were contributing. Moreover, the majority of the corpora are in English, suggesting that working with dialogues in other languages will require a greater effort.

Next section enumerates all the surveyed corpora and gives an overview of their analysis. After that, each section of this paper focuses on one specific aspect or a group of related aspects. Section 3 is on the speakers, corpus size and covered languages. Section 4 is on the type of linguistic annotations included in the corpora, also referring some of the tasks they can be used for. Section 5 is on the approach followed for collecting their data. Section 6 is on the contents of the corpora, namely their domain and related information. Before concluding, Section 7 builds on the previous section to discuss possible options that can be taken when a suitable corpus is not available.

2. Corpora Overview

This survey targets textual corpora of dialogue between human speakers, described in the literature, and created and used for research purposes. Alphabetically listed, the following corpora were included in this survey: AMI (Carletta et al., 2005); CamRest676 (Wen

et al., 2016; Wen et al., 2017); CoQA (Reddy et al., 2019); CrossWOZ (Zhu et al., 2020); DailyDialog (Li et al., 2017); DealOrNoDeal (Lewis et al., 2017); DECODA (Bechet et al., 2012); DIME (Villaseñor and Pineda, 2001); DSTC4 (Kim et al., 2017), DSTC5 (Kim et al., 2016); DSTC6, track 2 (Hori et al., 2019); DSTC7, tracks 1, 2 and 3 (D'Haro et al., 2020); EmotionLines (Hsu et al., 2018) and its extensions MELD (Poria et al., 2019) and EMOTyDa (Saha et al., 2020); ES-Port (García-Sardiña et al., 2018); Frames (El Asri et al., 2017); KVRET (Eric et al., 2017); MapTask (Anderson et al., 1991); Mastodon (Cerisara et al., 2018); MedDialog (Zeng et al., 2020); MRDA (Shriberg et al., 2004); MultiWOZ, covering version 2.0 (Budzianowski et al., 2018), as well as further corrections and new annotations introduced in versions 2.1 (Eric et al., 2020), 2.2 (Zang et al., 2020), 2.3 (Han et al., 2020) and 2.4 (Ye et al., 2021); OpenSubtitles (Tiedemann, 2009; Lison and Tiedemann, 2016); QuAC (Choi et al., 2018); Redial (Li et al., 2018); SAMsum (Gliwa et al., 2019); Schema-Guided Dialogue (SGD) (Rastogi et al., 2020); Switchboard (Godfrey et al., 1992) and SWDA (Jurafsky and Shriberg, 1997); Taskmaster-1 (Byrne et al., 2019); Topical-Chat (Gopalakrishnan et al., 2019); Ubuntu Dialog Corpus (UDC) (Lowe et al., 2015); and Wizard of Wikipedia (WOW) (Dinan et al., 2018).

Table 1 presents an overview of the results of this survey, with every corpora and their surveyed properties, namely: the number of speakers (where * stands for multiple or a variable number of speakers); the modalities they are available on (AUD for audio, TXT for text, VID for video); number of dialogues (when not available, alternative information is presented); language(s) covered (where * stands for multiple languages); linguistic annotations included; collection process; and domain.

The table shows right away that the surveyed corpora are significantly different in several aspects, from their size, to the covered domains. In the next sections we analyse some of these aspects.

3. Speakers, Size and Languages

The majority of these corpora is restricted to conversations between two human speakers, but some may include dialogues with more. This happens for AMI and MRDA, which have dialogues from meetings with several participants, or EmotionLines and OpenSubtitles, which have movie dialogues with a variable number of participants. SAMsum may also include dialogues with more than two speakers. On the other hand, UDC may include posts with no answer, and thus a single speaker.

Dialogue corpora are commonly used for training dialogue systems to respond as if they were humans, and so it makes sense that the dialogues they contain are indeed between humans. However, there are also corpora of conversations between humans and dialogue systems, typically including linguistic annotations, often made or revised by humans. These include corpora like ATIS (Price, 1990), DIHANA (Alcácer et al., 2005), LEGO (Schmitt et al., 2012), dialogues collected during the ConvAI challenges (Burtsev et al., 2018; Logacheva et al., 2020), and the corpora used in some editions of the Dialog State Tracking Challenge (DSTC) (Williams et al., 2013; Henderson et al., 2014a; Henderson et al., 2014b) or of its later rebranding as Dialog System Technology Challenge (Hori et al., 2019). The value of such corpora lies mainly in their annotations, useful for automating the process of dialogue analysis, which may later be useful for improving dialogue systems. For more on linguistic annotations in dialogue corpora, see section 4. Although we may refer to some of the previous corpora in the following sections, the main focus of this survey are corpora of dialogues exclusively between humans.

We find corpora of significantly variable sizes, ranging from just a few dozens (DIME, MapTask, MRDA, DSTC4, DSTC5) to thousands of dialogues. Of course, the size of the corpora is heavily constrained by the source of the data, domain covered and linguistic annotations made. Due to its nature, we do not have the number of dialogues in OpenSubtitles, but it is for sure the largest surveyed corpus, because it includes the subtitles for 152 thousand movies, in many languages, and each movie will have many dialogues. These dialogues are not annotated and there are many available on the web, produced by volunteers. Other large corpus are MedDialog, DSTC7-Track2 and UDC, respectively from a web forum, a social network, and chat logs, none of them with linguistic annotations. More on annotations, collection and domains can be found respectively in Sections 4, 5 and 6.

The majority of the surveyed corpora are in English, but there is a minority in other languages, namely French (DECODA), Spanish (DIME, ES-Port), or Chinese (CrossWOZ, MedDialog, DSTC5). The coverage of other languages is slightly broader if we consider human-machine dialogues, for which there are more corpora in Spanish (e.g., DIHANA) or in Japanese (e.g., DSTC6-Track3). Given its nature, OpenSubtitles covers a total of 60 languages and is available as parallel corpora. This means that researchers willing to work on different languages will probably have to find another corpus (e.g., private) or creating their own. In this case, the range of data collection approaches and domains covered by the surveyed corpora may serve as inspiration. For a discussion of available options, see section 7.

4. Linguistic Annotations

Most of the tasks in the scope of dialogue analysis benefit from having a dialogue corpus where computational models can be learned from. In some cases, annotations are not even necessary, as it happens for data-driven response generation, often framed as a problem of learning to translate user interactions to suitable re-

Name	Speakers	Modality	# Dialogues	Language	Annotations	Collection	Domain
AMI	4	AUD+TXT	100 hours	EN	DAs	recorded meetings	interacting w/ technology
CamRest676	2	TXT	676	EN	state	crowd (WOZ)	restaurants
CoQA	2	TXT	8,399	EN	question rationale	crowd (questioner-answerer)	7 domains
CrossWOZ	2	TXT	6,000	ZH	state	crowd (WOZ)	5 domains
DailyDialog	2	TXT	13,118	EN	DAs, emotion	web forum	daily communication
DealOrNoDeal	2	TXT	5,808	EN	items + values	crowdsourcing (chat)	bargaining
DECODA	2	AUD+TXT	1,514	FR	call type	recorded phone calls	urban transports
DIME	2	AUD+TXT	31	ES	DAs, discourse referents	crowd (WOZ)	kitchen design
DSTC4	2	TXT	35	EN	state	crowd (phone calls)	tourist information
DSTC5	2	TXT	70	EN, ZH	state	crowd (phone calls)	tourist information
DSTC6-Track2	2	TXT	1,024	EN	–	social network	airline, car, retail, fast food chains, etc.
DSTC7-Track1	2	TXT	135,893	EN	–	chat logs	advising, technical support
DSTC7-Track2	2	TXT	3M	EN	–	social network	open
DSTC7-Track3	2	TXT	7,659	EN	–	crowdsourcing (chat)	short videos
EmotionLines	*	AUD+VID+TXT	2,000	EN	DAs, emotion	movie subtitles	open
ES-Port	2	TXT	1,170	ES	acoustic events	calls to technical support	telecommunications
Frames	2	TXT	1,369	EN	frames	crowd (WOZ)	vacations
KVRET	2	TXT	3031	EN	slots	crowd (WOZ)	in-car personal assistant
MapTask	2	AUD+TXT	128	EN	behaviours	crowd (phone calls)	map instructions
Mastodon	2	TXT	505	EN	DAs, polarity	social network	open
MedDialog	2	TXT	3.6M	EN, ZH	–	web forum	health
MRDA	*	AUD+TXT	71	EN	DAs	recorded meetings	topics, debates, issues, social dynamics
MultiWOZ	2	TXT	10,000	EN	state	crowd (WOZ)	7 domains
OpenSubtitles	*	TXT	152k movies	*	–	movie subtitles	open
QuAC	2	TXT	13,594	EN	DAs	crowd (student-teacher)	people
Redial	2	TXT	10,000	EN	suggested, seen, liked	crowdsourcing (chat)	movie recommendations
SAMSum	$\geq$2	TXT	16,000	EN	summary	handcrafted by linguistics	messenger conversations
SGD	2	TXT	20,000	EN	DAs, state	crowd	20 domains
SWDA	2	AUD+TXT	1,155	EN	DAs	crowd (phone calls)	70 topics
Taskmaster-1	2	TXT	13,215	EN	API arguments	crowd	6 domains
Topical-Chat	2	TXT	11,000	EN	topic	crowd (chat)	8 topics
UDC	1-2	TXT	930,000	EN	–	chat logs	technical support
WOW	2	TXT	23,311	EN	topic	crowd (WOZ)	various topics

Table 1: Overview of dialogue corpora, where * stands for multiple values.

sponses. It is thus no surprise that machine translation approaches have been applied for generating responses to tweets (Ritter et al., 2011) and for modelling conversations (Vinyals and Le, 2015). Corpora without annotations may also be used in unsupervised approaches, e.g., clustering utterances according to the intent and discovering dialogue flows (Ritter et al., 2010).

However, whereas response generation systems do not always require a great understanding of the dialogue, when it comes to task-oriented dialogue systems, it is important to represent the current state of the dialogue at each turn. This comes as a set of linguistic annotations, which can be useful for training dialogue systems, and may also serve as common benchmarks, thus helping to evaluate progress in the task of their automatic prediction.

4.1. Dialogue State Tracking

Some corpora are annotated with useful information for a dialog-state architecture (Williams et al., 2016), where the meaning of the user's intents is represented as slots and their values (e.g., `from:downtown`, `inform:price=cheap`), grouped in a task-related frame where some of the values might be empty. The history of the dialogue may then be represented by a sequence of such frames.

Annotations for the previous tasks are included in the corpora of the first five editions of the DSTC, even if some are between humans and dialogue systems, but also for each customer utterance in MultiWOZ, CamRest676, CrossWOZ, and SGD, all including conversations between humans. KVRET adopts an alternative annotation, where the knowledge of the utterances is represented in the form of key-value pairs, directly converted to triples (e.g., `<dinner, time, 8pm>`). Towards a simpler annotation, Taskmaster restricts annotations according to the type of conversation, more precisely, to the variables required to execute the transaction, which the authors refer as API arguments. The Frames corpus goes beyond dialogue tracking and includes frame tracking annotations, which the authors claim to capture more complex dialogue flows, where it might be necessary to simultaneously keep track of several frames.

4.2. Dialogue Acts

Having in mind that each utterance in a dialogue is a kind of action performed by the speaker (Austin, 1962), roughly the speaker's intention, a simpler insight on each utterance is the dialogue act (DA), also known as speech act. Another common task in dialogue analysis is thus DA classification (DAC), which consists of labelling each utterance in a dialogue according to its DA. Given that the current DA often depends on the previous, DAC can be framed as a sequence labelling problem, instead of a simple classification task. It is thus no surprise that it has been attempted with Hidden Markov Models (Stolcke et al., 2000), Conditional Random Fields (Kim et al., 2010), or Recurrent Neural Networks with a LSTM layer (Kumar et al., 2018), among others.

There are domain-independent taxonomies of DAs (including, e.g., `greeting`, `question` or `acknowledge`), but the DA may also depend on the domain and on the task at hand (e.g., `request-address`, `inform-food`). On this context, the Dialog Act Markup in Several Layers (DAMSL) (Core and Allen, 1997) organises domain-independent DAs in four main categories (communicative, information, forward-looking, backward-looking) and has been frequently adopted. SWDA is one of the most popular corpus annotated with (an augmented version of) DAMSL, but other corpora adopt DAMSL or a subset of its tags, namely DIME, EMOTyDA, Mastodon and MRDA.

The utterances of DailyDialog are annotated according to four DA classes interoperable across multiple domains (Amanova et al., 2016), roughly matching the four main categories of DAMSL: inform, question, directive, commissive. Other corpora of human dialogues are annotated with typologies of DAs adapted to their tasks, namely AMI, MapTask, QuAC. There are also corpora of human-machine dialogues with annotated DAs, namely LEGO and DIHANA. The latter has the particularity of adopting a three-level hierarchy for DAs (speech act, data used / modified, specific data).

The DA is generally also included in datasets with dialogue tracking annotations (see e.g., MultiWOZ 2.1, CamRest676, SGD), which may thus be used for the task of DAC as well. However, while the dialogue state can be performed implicitly during the dialogue (e.g., in order to fulfil a given task, one of the human agents adds and fills slots in a form), the annotation of the DA is typically done at a later stage, and resorts to human annotators (Budzianowski et al., 2018).

4.3. Sentiment and Emotion

Due to their relevance for improving client-customer interactions, or predicting user opinions, another type of annotation in dialogue corpora is related to sentiment and emotion. In Mastodon, each utterance has an assigned polarity (positive, negative, neutral), and there are corpora where each utterance has an emotion label (DailyDialog, EmotionLines), corresponding to one of the six basic Ekman's emotions (Ekman, 1999) (joy, sadness, anger, fear, disgust, surprise), plus neutral. Given its correlation with DAs, sentiment classification in dialogues has been attempted jointly with DAC (Qin et al., 2020) and emotions have also proved to be useful for the latter task (Saha et al., 2021).

4.4. Other Annotations

Some corpora of spoken conversation also include annotations based on the actual speech and acoustic features, which may be related to emotion, but also cover behaviours like filled pauses, false starts and repetitions, broken words (MapTask, ES-Port); background

noise and laughter (ES-Port); head movements (AMI); or discourse referents (DIME). Still, since we are focusing on textual corpora, we did not look deep into these. Moreover, the transcriptions of the DECODA corpus have several semantic (call type) and syntactic annotations (disfluencies, parts-of-speech, chunks, named entities, syntactic dependencies).

We can say that the previous are the most common linguistic annotations we find in dialogue corpora. However, there are corpora with alternative annotations, such as summaries of the conversation (SAMsum), useful for developing abstractive summarisation tools; movies mentioned in a conversation and whether they had been suggested, seen or liked (Redial), useful, e.g., for conversational recommender systems (Jannach et al., 2021); the topic of the conversation (WOW, Topical-Chat), useful for developing systems capable of chatting on a close domain of knowledge; medical diagnosis and treatment suggestions (in some conversations of MedDialog); or the rationale for question-answering (QuAC).

We included in this survey two corpora with questions and answers in a conversation scenario (CoQA, QuAC). Each dialogue starts with a textual passage about which one of the speakers asks questions, while the other provides the answers, based on the aforementioned passage. These corpora are typically used for assessing approaches for extractive question-answering, which can be tackled by fine-tuning a neural language model like BERT (Devlin et al., 2019). However, they add the necessity of considering the conversation context for properly interpreting the questions.

5. Data Collection

One of the main goals of dialogue corpora is to provide training data for dialogue systems, so they can mimic human behaviours in certain tasks and scenarios. However, due to privacy reasons, only a minority of the surveyed dialogue corpora are collected from real situations (DECODA, ES-Por, MRDA). In the majority of the corpora, if not all, the speakers knew that their conversation was being recorded, which could, of course, condition their behaviour. But this is a necessary trade-off.

Several dialogue corpora are the result of spoken conversations and are thus available in the audio form. This is the case of some of the surveyed corpora (AMI, DECODA, DIME, MapTask, MRDA, Switchboard). However, for all the previous, manual or automatic transcriptions are also provided, meaning that they are in fact multimodal. In addition to audio and text, MELD and EMOTyDA add video to EmotionLines. Despite resulting from phone calls, due to the presence of sensitive data, only the (anonymised) transcription are available for ES-Port, not the audio. The same happens for DSTC4 and DSTC5.

More than knowing that the conversation was being recorded, in many corpora, the speakers were following a script and conversing with the single purpose of creating the corpus. Several of these follow the Wizard of Oz (WOZ) (Kelley, 1984) paradigm, where a conversation occurs between two speakers with different roles. One of them will play the role of a common user, also known as the questioner, to which a task is given (e.g., to find an entity, to chat about some topic). In order to fulfil this task, the questioner must interact, using natural language, with another user, the wizard, also known as the answerer, which will have access to more information on the domain (e.g., a database or a longer collection of documents) and possibly an interface that will help them track the user requests and efficiently provide suitable answers. This paradigm has been followed in the creation of spoken corpora (DIME), also including human-machine dialogues (ATIS, DIHANA), but mostly written conversation corpora (CrossWOZ, CamRest676, MultiWOZ, Frames, KVRET, WOW, Taskmaster). Other corpora created from crowdsourcing include DealOrNoDeal, Redial, Topical-Chat, CoQA and QuAC. All of the latter rely on the Amazon Mechanical Turk[1] platform, where tasks are prepared and workers are recruited to perform them in exchange of a monetary compensation.

Alternative sources to the WOZ paradigm include the exploitation of Web sources where users engage in conversations, including chat logs (UDC), forums (MedDialog), language learning websites (DailyDialog) and, of course, social networks. Due to the common presence of short-message dialogues, as well as its flexible API, Twitter has been used as the source of conversations (Ritter et al., 2011; Hori et al., 2019). However, due to its privacy-protecting restrictions, some researchers looked for similar but more relaxed sources, specifically Mastodon.

Movie subtitles, openly available in a large number in various websites, are another source of conversations. OpenSubtitles is a large corpus of this kind. Another example is the EmotionLines corpus, which is much smaller, focused on a single TV show (Friends), but with annotated emotions.

Approaches with simpler logistics were adopted for the creation of SAMSum, where conversations in the style of messenger apps were created and written down by linguists; part of Taskmaster, where dialogues were created by a single user, writing turns for both speakers, following a conversational scenario; or SGD, where conversations follow predefined flows, translated to natural language by crowd workers.

6. Domains & Contents

Humans can converse about virtually any topic, for their daily communication, regarding a specific situation, in order to achieve a predefined goal, or just for entertainment, with no specific purpose in mind.

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

For different topics and domains, a different vocabulary will be used, and the goal to achieve will constrain the conversation significantly. It is thus important to have corpora covering as much of such situations, topics and domains as possible, enabling different kinds of analysis and the development of dialogue systems with different purposes.

Despite the growing number of domains for which dialogue corpora are available, there will always be domains or specific applications, including different languages, for which such a corpus is nonexistent. In some cases, existing corpora may be still used or adapted, but in others new corpora will have to be created.

We identified several task-oriented corpora, namely those with dialogues between a customer and an agent. Many are focused on travelling or tourism-related domains, with customers asking for information such as locations, times, or suggestions that match their requirements. This includes vacations in general (Frames), restaurants (CamRest676), or urban transports (DECODA). When considering human-machine conversations, related domains include flights (ATIS), trains (DIHANA) and buses (LEGO). Another identified domain was in-car personal assistance (KVRET).

But some corpora cover more than one domain, often including but not restricted to tourism. Examples are restaurants, hotels, attractions, taxis, trains, hospitals and police (MultiWOZ); hotels, restaurants, attractions, metros, and taxis (CrossWOZ); hotels, flights, and car rentals (DSTC4, DTSC5); ordering pizza, creating auto repair appointments, setting up ride service, ordering movie tickets, ordering coffee drinks and making restaurant reservations (Taskmaster); or a range of 20 domains, including some of the previous, and others, such as alarm, calendar, events, media, messaging, movies or weather (SGD).

Though less restricted on the kind of possible requests, there are corpora roughly focused on other domains like customer-support in the technological (UDC, DSTC7-Track1) or telecommunications (ES-Port), as well as conversations between patients and doctors (MedDialog). The latter is especially interesting because health is a sensitive domain for which it is difficult to get data of this kind.

Some corpora cover tasks that involve the cooperation between two speakers. In DIME, the cooperative task is around kitchen design, and the speakers can refer to objects through a graphical user interface. In Map-Task, both speakers have a map, but only one has a route, which is to be instructed to the other to follow. In DealOrNoDeal, dialogues are on a multi-issue bargaining task, i.e., several items are available and human agents have to converse towards an agreement on the distribution of these items among them. In this process, agents try to maximise their reward function, different for each agent and not visible to the other.

Though not task-oriented, there are corpora with conversations on varied topics (WOW, Topical-Chat) or focused movies and recommendations (Redial); of questions and their answers (CoQA, QuAC); as well as of debates (MRDA, AMI) on different domains.

In addition to not being task-oriented, some corpora can be considered to be open domain. These include conversations from social networks (Mastodon, DSTC6-Track2, DSTC7-Track2); conversations in the style of those exchanged in messenger applications (SAMSum); daily life conversations in a website where English learners practice (DailyDialog); or TV show or movie subtitles (OpenSubtitles, EmotionLines).

7. Discussion

Despite the broad list of identified corpora, scenarios for which none of them applies can be easily envisioned. When this happens, an alternative would be to look at the available corpora and select the closest to the task at hand, possibly considering some adaptations. Otherwise, one will have to look at available sources of conversation data, get inspiration from the creation of the available dialogue corpora, and consider the creation of a new corpus.

In projects involving companies, it is normal that there is real data, which can and should be used, but, due to privacy legislation, its public release will generally not be possible. As we have shown in section 5, alternative sources of conversation data, available in many languages, include movie subtitles or web sources where people may leave their comments and establish conversations, like forums or social networks. Still, despite being publicly available through APIs or web scrapping, privacy laws might also apply and, in some cases, prevent the public distribution of the data as a corpus, even if for research purposes.

Another option would be to create all the data, from scratch. If resources are available, one may devise a crowdsourcing task, e.g., following the WOZ paradigm. Conversations would result from the interactions between paid workers, possibly following guidelines regarding the kind of dialogue to simulate. For task-oriented systems, to be used in customer-support, a speaker performing the role of customer / questioner would ask questions to the other, which would have access to more information, e.g., in the form of a database or a collection of textual documents. Possibly not as natural, but more practical, would be to adopt self-dialog, as in the Taskmaster corpus, i.e., a single user following a given scenario and writing all the turns of a dialogue. In any case, the previous approaches will condition the conversation, not desired for spontaneous speech, but acceptable for task-oriented dialogue.

As discussed in Section 4, for most dialogue analysis tasks, data alone is not enough, and linguistic annotations have to be made. Since these annotations are generally based on human theories, and they will be used for training and evaluating computational sys-

tems, most of the times they will have to be added manually. Depending on the available resources, as well as on the difficulty of explaining the task, they can be done by linguists, domain experts, or crowd workers. An alternative is to first produce the dialogue flows first, including all necessary annotations, and use them as the guidelines for a process like WOZ, i.e., ask humans to engage in a conversation that follows such flows. In the end, one could assume that the annotations would be valid for the resulting text.

Still regarding annotations, if the only problem is the language (e.g., there is a corpus for the target task, but it is not in the desired language), an option would be to translate the corpus automatically to the target language, while keeping the annotations. However, this is rarely a good option, due to potential issues on machine translation, including some specificities of each language that would hardly be captured, resulting in less natural text, and possibly incoherent annotations.

8. Conclusion

This was a brief survey on dialogue corpora available as text, with a focus on those between human speakers, that are publicly available or, according to the authors, can be provided for research purposes. We tried to cover a broad range of distinct corpora, including different languages, tasks, domains, and creation approaches, but were not so strict, so it is likely that we left out some corpora that matched our search criteria.

This survey may be useful for researchers working on projects involving the analysis of written dialogue, including, but not limited to, the development of dialogue systems. With this analysis, we could identify corpora of variable sizes and languages, created through different approaches, and covering different tasks and domains, but we could think of many more. The largest corpora have no annotations, whereas common annotations include dialogue acts and states, but in some cases also the sentiment tags. Although we focus on corpora available as text, some are originally spoken corpora that have been transcribed. Due to privacy constrains, most of the corpora were created in a crowdsourcing environment, where speakers are guided by assigned roles and tasks to accomplish.

Moreover, the majority of the corpora are in English, meaning that working with dialogues in other languages will generally require a greater effort, both on the collection of data and on its annotation. In most cases, this does not exactly mean that corpora is not available for other languages, only that the usage of such corpora is restricted and cannot be made publicly available. Consequently, there are no common benchmarks for those languages, making it harder to compare different works and to measure progress in the area. On top of this, interested researchers will possibly have to multiply the effort involved in data annotation.

Acknowledgements

This work was financially supported by the project FLOWANCE (POCI-01-0247-FEDER-047022), co-financed by the European Regional Development Fund (FEDER), through Portugal 2020 (PT2020), and by the Competitiveness and Internationalization Operational Programme (COMPETE 2020); and by national funds through FCT, within the scope of the project CISUC (UID/CEC/00326/2020) and by European Social Fund, through the Regional Operational Program Centro 2020.

9. Bibliographical References

Amanova, D., Petukhova, V., and Klakow, D. (2016). Creating annotated dialogue resources: Cross-domain dialogue act classification. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 111–117, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Austin, J. L. (1962). *How to do things with words: the William James lectures delivered at Harvard University in 1955*. Oxford University Press.

Burtsev, M., Logacheva, V., Malykh, V., Serban, I. V., Lowe, R., Prabhumoye, S., Black, A. W., Rudnicky, A., and Bengio, Y. (2018). The first conversational intelligence challenge. In *The NIPS'17 Competition: Building Intelligent Systems*, pages 25–46. Springer.

Core, M. G. and Allen, J. (1997). Coding dialogs with the DAMSL annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56, pages 28–35. Boston, MA.

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, 22(5):811–817.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Procs. of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. ACL Press.

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, pages 45–60.

Henderson, M., Thomson, B., and Williams, J. D. (2014a). The second Dialog State Tracking Challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272.

Henderson, M., Thomson, B., and Williams, J. D. (2014b). The third Dialog State Tracking Challenge. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 324–329. IEEE.

Jannach, D., Manzoor, A., Cai, W., and Chen, L. (2021). A survey on conversational recommender systems. *ACM Comput. Surv.*, 54(5), May.

Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.

Kim, S. N., Cavedon, L., and Baldwin, T. (2010). Classifying dialogue acts in one-on-one live chats. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 862–871.

Kumar, H., Agarwal, A., Dasgupta, R., and Joshi, S. (2018). Dialogue act sequence labeling using hierarchical encoder with CRF. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Logacheva, V., Malykh, V., Litinsky, A., and Burtsev, M. (2020). Convai2 dataset of non-goal-oriented human-to-bot dialogues. In *The NeurIPS'18 Competition*, pages 277–294. Springer.

Qin, L., Che, W., Li, Y., Ni, M., and Liu, T. (2020). Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8665–8672.

Ritter, A., Cherry, C., and Dolan, W. B. (2010). Unsupervised modeling of Twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180.

Ritter, A., Cherry, C., and Dolan, W. B. (2011). Data-driven response generation in social media. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pages 583–593. Association for Computational Linguistics.

Saha, T., Gupta, D., Saha, S., and Bhattacharyya, P. (2021). Emotion aided dialogue act classification for task-independent conversations in a multi-modal framework. *Cognitive Computation*, 13(2):277–289.

Serban, I. V., Lowe, R., Henderson, P., Charlin, L., and Pineau, J. (2018). A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue & Discourse*, 9(1):1–49.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

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

Williams, J., Raux, A., Ramachandran, D., and Black, A. (2013). The Dialog State Tracking Challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413, Metz, France, August. Association for Computational Linguistics.

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

10. Language Resource References

Alcácer, N., Benedı, J., Blat, F., Granell, R., Martınez, C., and Torres, F. (2005). Acquisition and labelling of a spontaneous speech dialogue corpus. In *Proceeding of 10th International Conference on Speech and Computer (SPECOM). Patras, Greece*, pages 583–586.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The HCRC map task corpus. *Language and speech*, 34(4):351–366.

Bechet, F., Maza, B., Bigouroux, N., Bazillon, T., El-Bèze, M., De Mori, R., and Arbillot, E. (2012). DECODA: a call-centre human-human spoken conversation corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1343–1347, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gasic, M. (2018). MultiWOZ – a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Byrne, B., Krishnamoorthi, K., Sankar, C., Neelakantan, A., Goodrich, B., Duckworth, D., Yavuz, S., Dubey, A., Kim, K.-Y., and Cedilnik, A. (2019). Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., et al. (2005). The AMI meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Cerisara, C., Jafaritazehjani, S., Oluokun, A., and Le, H. T. (2018). Multi-task dialog act and sentiment recognition on Mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 745–754.

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

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. (2018). Wizard of Wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

D'Haro, L. F., Yoshino, K., Hori, C., Marks, T. K.,

Polymenakos, L., Kummerfeld, J. K., Galley, M., and Gao, X. (2020). Overview of the seventh dialog system technology challenge: DSTC7. *Computer Speech & Language*, 62:101068.

El Asri, L., Schulz, H., Sarma, S. K., Zumer, J., Harris, J., Fine, E., Mehrotra, R., and Suleman, K. (2017). Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219.

Eric, M., Krishnan, L., Charette, F., and Manning, C. D. (2017). Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49.

Eric, M., Goel, R., Paul, S., Sethi, A., Agarwal, S., Gao, S., Kumar, A., Goyal, A., Ku, P., and Hakkani-Tur, D. (2020). MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France, May. European Language Resources Association.

García-Sardiña, L., Serras, M., and del Pozo, A. (2018). ES-port: a spontaneous spoken human-human technical support corpus for dialogue research in Spanish. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Gliwa, B., Mochol, I., Biesek, M., and Wawer, A. (2019). SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, November. Association for Computational Linguistics.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R., and Hakkani-Tür, D. (2019). Topical-Chat: Towards knowledge-grounded open-domain conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Han, T., Liu, X., Takanobu, R., Lian, Y., Huang, C., Peng, W., and Huang, M. (2020). MultiWOZ 2.3: A multi-domain task-oriented dataset enhanced with annotation corrections and co-reference annotation. *arXiv preprint arXiv:2010.05594*.

Hori, C., Perez, J., Higashinaka, R., Hori, T., Boureau, Y.-L., Inaba, M., Tsunomori, Y., Takahashi, T., Yoshino, K., and Kim, S. (2019). Overview of the sixth dialog system technology challenge: DSTC6. *Computer Speech & Language*, 55:1–25.

Hsu, C.-C., Chen, S.-Y., Kuo, C.-C., Huang, T.-H., and Ku, L.-W. (2018). EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Jurafsky, D. and Shriberg, E. (1997). Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Technical report, University of Colorado at Boulder &+ SRI International.

Kim, S., D'Haro, L. F., Banchs, R. E., Williams, J. D., Henderson, M., and Yoshino, K. (2016). The fifth Dialog State Tracking Challenge. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 511–517. IEEE.

Kim, S., D'Haro, L. F., Banchs, R. E., Williams, J. D., and Henderson, M. (2017). The fourth Dialog State Tracking Challenge. In *Dialogues with Social Robots*, pages 435–449. Springer.

Lewis, M., Yarats, D., Dauphin, Y., Parikh, D., and Batra, D. (2017). Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Li, R., Kahou, S. E., Schulz, H., Michalski, V., Charlin, L., and Pal, C. (2018). Towards deep conversational recommendations. In Samy Bengio, et al., editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9748–9758.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May. European Language Resources Association (ELRA).

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

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Price, P. (1990). Evaluation of spoken language sys-

tems: The ATIS domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Rastogi, A., Zang, X., Sunkara, S., Gupta, R., and Khaitan, P. (2020). Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

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

Saha, T., Patra, A., Saha, S., and Bhattacharyya, P. (2020). Towards emotion-aided multi-modal dialogue act classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4361–4372.

Schmitt, A., Ultes, S., and Minker, W. (2012). A parameterized and annotated spoken dialog corpus of the CMU Let's Go bus information system. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3369–3373, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. (2004). The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100.

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

Villaseñor, Luis, A. M. and Pineda, L. A. (2001). The DIME corpus. In *Encuentro Internacional de Ciencias de la Computación, vol. 2, 1–10*.

Wen, T.-H., Gasic, M., Mrkšić, N., Barahona, L. M. R., Su, P.-H., Ultes, S., Vandyke, D., and Young, S. (2016). Conditional generation and snapshot learning in neural dialogue systems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2153–2162.

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

Ye, F., Manotumruksa, J., and Yilmaz, E. (2021). Multiwoz 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. *arXiv preprint arXiv:2104.00773*.

Zang, X., Rastogi, A., Sunkara, S., Gupta, R., Zhang, J., and Chen, J. (2020). MultiWOZ 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online, July. Association for Computational Linguistics.

Zeng, G., Yang, W., Ju, Z., Yang, Y., Wang, S., Zhang, R., Zhou, M., Zeng, J., Dong, X., Zhang, R., et al. (2020). MedDialog: A large-scale medical dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250.

Zhu, Q., Huang, K., Zhang, Z., Zhu, X., and Huang, M. (2020). CrossWOZ: A large-scale Chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295.

A. Corpora URLs

Table 2 compiles the URLs where the surveyed corpora can be downloaded from, as of December 2021. Out of them, DECODA is not available because, according to its authors, includes much personal data, hard to anonymise. Moreover, DSTC4 and DSTC5 have restricted availability, exclusive to the participants in the challenges where they were used. For ES-Port, we found two download links, but none was working in April 2022.

Name	URL
AMI	https://groups.inf.ed.ac.uk/ami/download/
CamRest676	https://github.com/WING-NUS/sequicity/tree/master/data/CamRest676
CoQA	https://stanfordnlp.github.io/coqa/
CrossWOZ	https://github.com/thu-coai/CrossWOZ
DealOrNoDeal	https://github.com/facebookresearch/end-to-end-negotiator
DailyDialog	http://yanran.li/dailydialog.html
DECODA	N/A
DIME	https://data.stanford.edu/dime
DSTC4	https://colips.org/workshop/dstc4/ (restricted to participants in the challenge)
DSTC5	https://github.com/seokhwankim/dstc5 (restricted to participants in the challenge)
DSTC6	https://github.com/dialogtekgeek/DSTC6-End-to-End-Conversation-Modeling
DSTC7	https://github.com/mgalley/DSTC7-End-to-End-Conversation-Modeling
EmotionLines	http://doraemon.iis.sinica.edu.tw/emotionlines/index.html
ES-Port	https://aholab.ehu.eus/metashare/repository/browse/es-port-the-spanish-technical-support-corpus/b5643034100f11e8b066f01faff11afaa7fc3195f8a64c929bb313140a170cdb/ (link not working)
Frames	https://www.microsoft.com/en-us/research/project/frames-dataset/
KVRET	https://metatext.io/datasets/a-multi-turn,-multi-domain-dialogue-dataset-(kvret)
Mastodon	https://github.com/cerisara/DialogSentimentMastodon
MedDialog	https://github.com/UCSD-AI4H/Medical-Dialogue-System
MRDA	https://github.com/NathanDuran/MRDA-Corpus
MapTask	https://groups.inf.ed.ac.uk/maptask/
MultiWOZ	https://github.com/budzianowski/multiwoz
OpenSubtitles	https://opus.nlpl.eu/OpenSubtitles-v2018.php
QuAC	http://quac.ai/
Redial	https://redialdata.github.io/website/
SAMsum	https://arxiv.org/abs/1911.12237
SGD	https://github.com/google-research-datasets/dstc8-schema-guided-dialogue
SWDA	http://compprag.christopherpotts.net/swda.html
Taskmaster-1	https://research.google/tools/datasets/taskmaster-1/
Topical-Chat	https://github.com/alexa/Topical-Chat
UDC	https://github.com/rkadlec/ubuntu-ranking-dataset-creator
WOW	https://parl.ai/projects/wizard_of_wikipedia/

Table 2: URLs of the dialogue datasets, as of December 2021.

ANEXO D - COAUTORIA ARTIGO PUBLICADO - RECPAD 22

Natural Language Processing in the Classification of Dialog Act

Patrícia Ferreira[12]
patriciaf@dei.uc.pt
Daniel Martins[3]
daniel.martins2997@gmail.com
Hugo Gonçalo Oliveira[12]
hroliv@dei.uc.pt
Catarina Silva[12]
catarina@dei.uc.pt
Ana Alves[13]
ana@dei.uc.pt

[1] CISUC, Universidade de Coimbra, Portugal
[2] DEI, FCTUC, Universidade de Coimbra, Portugal
[3] ISEC, Instituto Politécnico de Coimbra, Portugal

Abstract

Dialog act (DA) classification is generally done on spontaneous utterances of dialog, and this paper aims to understand whether generalization through its annotations leads to improvements in DA classification. For this, Natural Language Processing (NLP) tasks were explored, such as, Part-of-Speech (PoS) Tagging, Named Entity Recognition (NER) and coreference resolution in the context of dialog. These tasks were implemented using the spaCy library and the result of them can be used as resources for the DA classification. Two datasets were used for these experiments: CamRest676 and Mastodon, presenting interesting insights into the possibility of using NLP in DA classification.

1 Introduction

Dialog act classification (DAC) is the task of classifying an utterance with respect to the function it serves in a dialog, i.e. the act the speaker is performing. Dialog acts are a type of speech acts [4].

Classification of dialog acts is not normally done on generalized dialogs, that is, on information that represents a class of DA. This article presents three techniques that extract information that allow the creation of a model that generalizes and allows improving the DA classification.

Hence, the main objective of this work is to understand if this generalization, using NLP tasks, can improve DA classification. For this, tasks of an NLP tool, spaCy [1], applied to dialogs were explored. Additionally, relationship extraction was carried out, as well as other lower-level NLP tasks that can be useful for text analysis, for example extracting keywords, entities and dialog relationships. The best results for DA classification depending on the application of these approaches are reported. These tasks were applied to dialogs from two corpora namely CamRest676 and Mastodon. To the best of our knowledge, none of these NLP tasks has been applied in DA classification.

2 Proposed Approach

The tasks of PoS Tagging, NER and Coreference Resolution were applied in the classification of DAs, in order to understand if they have an impact on it. We use spaCy v3.2 (February 2022).

2.1 NLP Tools and Corpora

NLP toolkits cover a set of tasks, starting with the traditional pipeline, e.g., tokenization, PoS tagging, NER and other types of chunking. For this, spaCy was used, which is a Python NLP toolkit that aims at performance over time and includes several pipelines for different languages.

A selection was made from several currently available dialog corpora, evaluating various aspects such as speakers, size, languages, collection, annotations and domains [3]. Two datasets publicly available in conversational textual formats were used for these experiments, namely CamRest676 [5] and Mastodon [1]. CamRest676 has task-oriented dialogs between a user, asking for restaurants with specific attributes, and an assistant that attempts to respond to the user's requests. We used CamRest676 already split into train, validation, and test, with 406, 135, and 135 dialogs respectively, giving a total of 676 dialogs, as the name implies. Mastodon is an open domain conversation dataset with social network conversations, and is similar to Twitter in terms of content and usage. Mastodon consists of 592 dialogs: 303 for testing and 289 for training.

[1] https://spacy.io/

2.2 Part-of-Speech Tagging

One of the explored tasks in preprocessing and text analysis was PoS Tagging, because some PoS tend to be more relevant than others, and PoS tagging may help to generalize utterances. Table 1 shows the result of applying PoS Tagging to a CamRest utterance, where PoS corresponds to the simple universal PoS tags, while the Tag is more detailed.

Word	PoS	Tag	Definition
There	PRON	EX	existential there
are	VERB	VBP	verb, non-3rd person singular present
several	ADJ	JJ	adjective, other noun-modifier
restaurants	NOUN	NNS	noun, plural

Table 1: PoS Tagging - CamRest676

In this work, PoS Tagging was applied using only the verb phrase, made up of main verbs, auxiliary verbs and objects, in order to understand the influence it could have on DA classification.

2.3 Named Entity Recognition (NER)

NER aims to identify and classify key information in text. More precisely, it delimits mentions of entities and classifies them accordingly. Like PoS Tagging, it is also useful for generalizing statements and allows understanding text data and probably obtaining more accurate results.

Figure 1 illustrates the application of the spaCy English NER module to CamRest676 assertions, where we can see that words are divided and classified according to their category, which can be a place, person, organization, time, object, or geographic entity. It is important to identify and classify the entities so that the model that will work on the data can easily understand the text data.

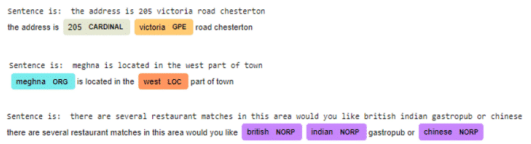


Figure 1: Viewing NER - CamRest676

2.4 Coreference Resolution

One challenge in NLP is that different names might be used for the same entity (e.g., Artificial Intelligence and AI) and, in some cases, pronouns are used (e.g., The Professor gave us a task. He was not happy.). Different expressions that refer to the same entity are known as coreferences. Thus, coreference resolution is the task of finding and handling such expressions and it is an important step for several higher-level NLP tasks that involve natural language understanding, such as document summarization, question answering and dialog analysis.

The standard spaCy pipeline does not have a module for coreference resolution, but a module, neuralcoref, can be installed for this purpose.

2.5 Application to Dialog Act Classification

We have shown what we can get from a dialog when applying different NLP tasks. Extracted information can be used as features for describing groups of related utterances, later useful for discovering dialog flows.

A possible abstract representation of such flows are sequences of DAs. Each utterance in a dialog is a kind of action performed by the speaker, and the DA labels this action (e.g., inform, request, etc.).

DA Classification is the task of automatically classifying the DA of an utterance, we did DA Classification, in both datasets, using traditional machine learning approaches. To see if DA Classification can take any advantage of the approaches above, the impact they can have on the preprocessing and classification phase was analyzed. For Mastodon the 8 original classes were used (Inform, suggestion, question, greetings, request, thanking, exclamation, and other). The original CamRest676 dataset is specified with domain information which results in 16 classes (e.g. inform-food, inform-name, inform-area, etc). We experimented with the 16 classes and with only four, that is, CamRest676 was used with the utterance labeled with one or more than three possible DAs (Inform, Request, Nooffer or None - if the utterance does not correspond to the DAs). As the conclusions were similar, we focused only on the results for the four DAs.

Two different classifiers were tested for labeling the utterances, namely: a Support Vector Classifier (SVC) and a Random Forest classifier (RF). For both, we adopted a One-vs-the-Rest (OvR) strategy. The classifiers were trained in the utterances of the train (and validation for CamRest676) portions and evaluated in the test portion.

Frequent word vectors wont be sparse, though the word may not be important. Rare words looks very sparse, hence less importance. To tackle these problems, we analyzed DAC with the utterances represented with TF-IDF (Term Frequency-Inverse Document Frequency) vectors, with max-df = 0.8, ignores terms that appear in more than 80% of the documents; min-df = 5, ignores terms that appear in less than 5 documents.

In order to analyze the PoS Tagging impact (considering verbs), NER and Coreference Resolution, we created five versions of the dataset namely:

- Original data (Original) - CamRest676 [5] and Mastodon [1]. Number of features for CamRest676 - 506; for Mastodon - 672;
- PoS Tagging identifying the verb phrase (Verb phrase). Number of features for CamRest676 - 188; for Mastodon - 274;
- Named Entities replaced by their category (NER). Number of features for CamRest676 - 386; for Mastodon - 641;
- Pronouns replaced by their referring expression (CR). Number of features for CamRest676 - 505; for Mastodon - 669;
- Named Entities replaced by their category and pronouns replaced by their referring expression (NER+CR). Number of features for CamRest676 - 413; for Mastodon - 639.

3 Results and Discussion

Tables 2 and 3 show the results of DA Classification, respectively in CamRest and in Mastodon.

Three distinct metrics were calculated, namely: Precision (P) - the proportion of positives that is actually correct; Recall (R) - the proportion of actual positives that was identified correctly and F1-Score (F1) - the harmonic mean of precision and recall. We also analyzed two averages for better understanding the global results: the micro-average, which considers each class individually; and the macro-average, which considers the classes as a whole, better reflecting the statistics of smaller classes [2].

		Original			W/ verb phrase			NER			CR			NER+CR		
	DA	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
S V C	Inform	0.95	**0.98**	0.96	**0.97**	**0.98**	**0.97**	0.96	**0.98**	**0.97**	0.95	0.97	0.96	0.95	0.97	0.96
	Request	0.95	**0.92**	0.93	**0.96**	**0.92**	**0.94**	0.95	**0.92**	**0.94**	0.95	0.90	0.93	0.93	0.89	0.91
	Nooffer	0.95	0.82	0.88	**0.99**	**0.86**	**0.92**	0.94	0.84	0.89	0.95	0.80	0.87	0.97	**0.86**	0.91
	None	**1.00**	0.91	0.95	0.99	**0.95**	**0.96**	**1.00**	0.92	**0.96**	**1.00**	0.91	0.95	0.98	0.91	0.95
	Micro-avg	0.96	0.94	0.95	**0.97**	**0.95**	**0.96**	**0.97**	0.94	0.95	0.96	0.93	0.95	0.95	0.93	0.94
	Macro-avg	0.96	0.90	0.93	**0.98**	0.92	**0.95**	0.96	0.92	**0.95**	0.96	0.90	0.93	0.93	**0.94**	0.93
R F	Inform	0.94	**0.97**	0.95	**0.95**	**0.97**	**0.96**	**0.95**	**0.97**	**0.96**	0.94	**0.97**	0.95	**0.95**	**0.97**	**0.96**
	Request	0.94	0.88	0.91	**0.95**	**0.92**	**0.93**	0.94	0.88	0.91	0.93	0.88	0.90	0.94	0.86	0.90
	Nooffer	0.97	0.67	0.79	**1.00**	**0.75**	**0.85**	0.98	0.71	0.82	0.97	0.63	0.76	0.98	0.65	0.78
	None	**0.99**	0.91	0.95	0.98	**0.93**	**0.96**	**0.99**	**0.93**	**0.96**	**0.99**	0.91	0.95	**0.99**	0.92	0.95
	Micro-avg	0.95	0.92	0.94	**0.96**	**0.94**	**0.95**	**0.96**	0.93	0.94	0.95	0.92	0.94	**0.96**	0.92	0.94
	Macro-avg	0.96	0.86	0.90	**0.97**	0.89	0.93	**0.97**	0.87	0.91	0.95	**0.92**	**0.95**	0.96	0.85	0.90

Table 2: DA Classification with SVC and RF on CamRest676 dataset

		Original			W/ verb phrase			NER			CR			NER+CR		
	DA	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
S V C	Inform	0.77	0.97	0.86	0.77	0.97	0.86	0.77	0.97	0.86	0.77	0.97	0.86	0.77	0.97	0.86
	Suggestion	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Question	**0.83**	0.21	0.34	0.74	**0.27**	**0.40**	0.80	0.18	0.30	0.82	0.20	0.32	0.78	0.17	0.28
	Greetings	0.80	0.47	0.59	0.73	0.47	0.57	**0.90**	**0.53**	**0.67**	0.80	0.47	0.59	0.80	0.47	0.59
	Thanking	**1.00**	0.81	0.90	0.93	**0.88**	0.90	**1.00**	**0.88**	**0.93**	**1.00**	0.81	0.90	**1.00**	**0.88**	**0.93**
	Other	**1.00**	0.07	0.13	0.75	0.07	0.13	**1.00**	0.07	0.13	1.00	0.07	0.13	**1.00**	0.07	0.13
	Exclamation	1.00	0.09	0.16	1.00	0.09	0.16	1.00	0.09	0.16	1.00	0.09	0.16	1.00	0.09	0.16
	Request	0.91	0.20	0.32	0.88	0.14	0.24	**1.00**	**0.21**	**0.33**	0.91	0.20	0.32	0.91	0.20	0.32
	Micro-avg	0.77	0.73	0.75	0.77	**0.74**	0.75	**0.78**	**0.74**	**0.76**	**0.78**	0.73	0.75	0.77	0.73	0.75
	Macro-avg	0.79	0.35	0.41	0.72	**0.36**	0.41	**0.80**	**0.36**	**0.42**	0.79	0.35	0.41	0.79	**0.36**	0.41
R F	Inform	0.79	0.93	0.86	0.79	0.93	0.86	**0.80**	**0.94**	0.86	**0.80**	**0.94**	0.86	0.79	**0.94**	0.86
	Suggestion	**0.67**	0.08	0.14	0.25	**0.12**	**0.16**	**0.67**	0.08	0.14	0.50	0.08	0.14	**0.67**	0.08	0.14
	Question	0.84	0.32	0.47	0.70	0.33	0.45	**0.86**	**0.34**	**0.49**	0.84	0.31	0.45	0.81	**0.34**	0.48
	Greetings	0.63	0.53	0.58	**0.71**	**0.59**	**0.65**	0.64	0.53	0.58	0.60	0.53	0.56	0.60	0.53	0.56
	Thanking	**0.94**	**0.94**	**0.94**	0.88	0.88	0.88	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**
	Other	0.67	0.14	0.23	**0.72**	**0.30**	**0.43**	0.70	0.16	0.26	0.70	**0.16**	0.26	0.70	0.16	0.26
	Exclamation	0.75	0.13	0.22	**0.83**	**0.22**	**0.34**	0.75	0.13	0.22	0.75	0.13	0.22	0.75	0.13	0.22
	Request	0.82	0.27	0.41	0.73	**0.31**	**0.44**	**0.88**	0.27	0.42	0.81	0.25	0.39	0.82	0.27	0.41
	Micro-avg	0.77	0.73	0.75	0.78	**0.75**	0.76	**0.80**	**0.75**	**0.78**	**0.80**	0.74	0.77	0.79	0.74	0.77
	Macro-avg	0.77	0.42	0.48	0.70	**0.46**	**0.52**	**0.78**	0.42	0.49	0.74	0.42	0.48	0.77	0.42	0.49

Table 3: DA Classification with SVC and RF on Mastodon dataset

Analyzing the metrics in the previous tables, it is verified that the best performance was obtained with a traditional Supervised Machine Learning for the 4 classes of CamRest676, with the best MicroF1 and MacroF1 being 95%, with variable performance between classes, lower for the least represented. As for Mastodon, Micro and Macro F1 were lower with best values of 78% and 49%, respectively. The results of these metrics are a little higher when NER is applied, but the general performances are very close, with no significant difference between values, suggesting that there is no real impact of the application from NER and/or CR.

As for the results of the classification of acts of dialog when only the verb phrase is applied, it is possible to verify that, in most cases, especially for CamRest676, excellent DAC values were obtained. Another advantage of applying TF-IDF with verbs is that the number of features is much lower compared to other approaches.

4 Conclusions and Future Work

In this paper, spaCy tools were explored for dialog data processing, covering tasks for linguistic processing, namely PoS tagging, useful for acquiring more abstract characteristics of utterances in a dialog; CR, which allows you to rewrite the text in a more user-friendly way to high-level tasks; and NER that locates and classifies named entities in the text into predefined categories. The output of the explored tasks could be useful to better describe the groups of related utterances and for more generalizable representations of utterances, for better performance on tasks like DA Classification. These approaches were applied to DA Classification. Interesting results were achieved, but the application of NER and CR did not make much difference.

In the future, more preprocessing approaches will be explored and other DAC experiments will be designed, considering other resources (e.g., speaker, turn number, etc.) and other representations of utterances.

Acknowledgements This work was financially supported by the project FLOWANCE (POCI-01-0247-FEDER-047022), cofinanced by FEDER , through PT2020 , and by COMPETE 2020 ; and by national funds through FCT, within the scope of the project CISUC (UID/CEC/00326/2020) and by European Social Fund, through the Regional Operational Program Centro 2020.

References

[1] Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa Le. Multi-task Dialog Act and Sentiment Recognition on Mastodon. *arXiv preprint arXiv:1807.05013*, 2018.

[2] Nadia Ghamrawi and Andrew McCallum. Collective multi-label classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 195–200, 2005.

[3] Hugo Gonçalo Oliveira, Patrícia Ferreira, Daniel Martins, Catarina Silva, and Ana Alves. A Brief Survey on Textual Dialogue Corpora. *Proc. of LREC*, 2022.

[4] Norbert Reithinger and Martin Klesen. Dialogue act classification using language models. In *Fifth EUROSPEECH*, 1997.

[5] Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*, 2016.

ANEXO E - TABELAS COM A APLICAÇÃO DAS DIFERENTES TÉCNICAS DE ETIQUETAGEM A TODOS OS CORPORA

Etiquetagem no CamRest676 de 4 *clusters*

Cluster	Bigrama	Sintagma Verbal	Fala mais próxima do centroide
***Cluster* 1**	thank you	thank you goodbye	thank you, goodbye.
***Cluster* 2**	in the	look be for	there are many expensive restaurants in town....
***Cluster* 3**	phone number	be what	what is the address and phone number?
***Cluster* 4**	thank you 1	care do not	you are welcome.

Etiquetagem no Mastodon

Cluster	Bigrama	Sintagma Verbal	Fala mais próxima do centroide
***Cluster* 1**	of the	be should not hard	that right ?
***Cluster* 2**	you re	feel do ever anxious at	i was really trying to sort out what a good re...
***Cluster* 3**	is the	say system do not word window mobile duck fu...	it is . until some update inevitably breaks ev...
***Cluster* 4**	ðÿ ðÿ	ugly	ur on mastodon
***Cluster* 5**	social media	delete account ago feel and never free	they are as tired of social media as i am .
***Cluster* 6**	of the 1	be probably count	my favorite toku themes are on my home timelin...
***Cluster* 7**	thank you	thank you	thank you

Etiquetagem no DailyDialog

Cluster	Bigrama	Sintagma Verbal	Fala mais próxima do centroide
***Cluster* 1**	Do you	Be waht	i know only too well what that's like
***Cluster* 2**	Yes it	Yes	yes, in most circumstances.
***Cluster* 3**	Ok ll	Ok	ok. how about some lamb kebabs?
***Cluster* 4**	It really	Really	it really hurts

Etiquetagem no CamRest676 de 16 *clusters*

Cluster	Bigrama	Sintagma Verbal	Fala mais próxima do centroide
***Cluster* 1**	phone number	be be	the phone number is 01223 327908, and the add...
***Cluster* 2**	good bye	thank you good bye	thank you good bye.
***Cluster* 3**	in the	be restaurant	there are mexican, italian, chinese and india...
***Cluster* 4**	phone number	be what	what is the address and phone number?
***Cluster* 5**	how about	about how food	what kind of food are you looking for?
***Cluster* 6**	of town	locate it be in	which part of town it is in ?
***Cluster* 7**	is there	be there anything	is there anything else?
***Cluster* 8**	thank you	thank no you	you're very welcome.
***Cluster* 9**	in the 1	be there restaurant	could you please find me a different restaura...
***Cluster* 10**	price range	be what 1	what price range do you prefer ?
***Cluster* 11**	thank you 1	thank you goodbye	thank you, goodbye.
***Cluster* 12**	in the 2	be restaurant 1	i would like a moderately priced chinese rest...
***Cluster* 13**	thank you 2	thank you	thank you.
***Cluster* 14**	don care	care do not	i don't care.
***Cluster* 15**	thank you 3	thank you for goodbye	thank you for using our system. good bye
***Cluster* 16**	in the 3	look be for	i'm looking for a moderately priced restaurant...

Etiquetagem no MultiWOZ *intents*

Cluster	Bigrama	Sintagma Verbal	Fala mais próxima do centroide
***Cluster* 1**		like would	
	to leave	leave	great no. that was it. thank you.
***Cluster* 2**			i'm looking for a place to stay in the
	need train	need train	north maybe a guesthouse.
***Cluster* 3**			i need to reserve a table at shiraz, can
	in the	look be for	you help me?
***Cluster* 4**			yes i am looking for someplace to go in
	thank you	thank you	the south for entertainment.
***Cluster* 5**	phone		i need train reservations from norwich
	number	be what	to cambridge
***Cluster* 6**	in the 1	look be for 1	cheap and in the south please.
***Cluster* 7**			hello. i need train to london liverpool
	price range	be what 1	street.
***Cluster* 8**		like would	i do not need a reservation. i need a
	for people	book	postcode and address only.
***Cluster* 9**			could you recommend an expensive
	free parking	look for	restaurant in the same area?
***Cluster* 10**			thanks. i will also need a taxi from the
	to		hotel to the restaurant. will you handle
	Cambridge	look be for 2	this?
***Cluster* 11**			before booking, i would also like to know the travel time, price, and
	taxi to	need book	departure time please.
***Cluster* 12**			alright got you booked on train tr4987, the total fee is 70.8 gbp payable at the station. your reference number is :
		like would I	9uprwuhb . anything else i can help with
	you like	book	today?
***Cluster* 13**			great, the tr1992 leaves at 21:35. would
	you like 1	be there train	that work for you?
***Cluster* 14**		complete	i have 133 trains matching your request.
	reference	type book	is there a specific day and time you
	number	car number	would like to travel?
***Cluster* 15**			i can help you with that. where are you
	in the	be hotel	going?
***Cluster* 16**			i have the cambridge museum of technology located in the east side of town. it's at the site of the old pumping station on cheddar's lane. would you like
	have great	have day	more information?

***Cluster* 17**			yes i have many options. what is your preferred price point?
	number is	be	
***Cluster* 18**		be there	i can help you with that. what day will you be traveling?
	in the 1	restaurant	
***Cluster* 19**			i recommend tenpin it's in cambridge leisure park, clifton way postcode cb17dy and their number is 08715501010. is there anything else i can help with today?
	Cambridge towninfo	thank you for have day	
***Cluster* 20**			we have many great places to stay. what area would you like and do you have a price range?
	like to	be what	
***Cluster* 21**			the top 3 places serving indian cuisine are curry garden, the golden curry, and saffron brasserie. would you like more options?
	anything else	be there anything	
***Cluster* 22**	in the 2	be in	thank you for using our service!

Etiquetagem no SNIPS

Cluster	Bigrama	Sintagma Verbal	Fala mais próxima do centroide
***Cluster* 1**	movie schedule	be what	show me the movie schedule for nearby movies
***Cluster* 2**	add artist	add tune to	add this song to my esenciales playlist
***Cluster* 3**	weather forecast	be what_1	what is the weather forecast here
***Cluster* 4**	tv series	find show	i d like to find the days of glory saga
***Cluster* 5**	rate current	give out to	rate this book 0 of 6 points
***Cluster* 6**	book table	book restaurant	i want to book a restaurant
***Cluster* 7**	play music	play music from	play the what

Etiquetagem no SGD de atos de diálogo

Cluster	Bigrama	Sintagma Verbal	Fala mais próxima do centroide
***Cluster* 1**			yes i am looking for someplace to go
	in the	look be for	in the south for entertainment.
***Cluster* 2**		like would	yes, book me a table for 2 people at
	for people	book	12:15 on monday.
***Cluster* 3**			yes it does. can you book the train for
	the train	need train	8 people?
***Cluster* 4**	price range	be what	does it have a star rating of 2?
***Cluster* 5**	thank you	thank you	thanks and i look forward to my stay
***Cluster* 6**	to		i need train reservations from norwich
	cambridge	look be for 1	to cambridge
***Cluster* 7**		like would	i would like chinese. i would like that
	to leave	leave	for 7pm.
***Cluster* 8**			could you recommend an expensive
	in the 1	look be for 2	restaurant in the same area?
***Cluster* 9**	thank you 1	thank you for	great, thanks for all of your help.
***Cluster* 10**	entrance fee	be what 1	yes, please book 5 seats for me.
***Cluster* 11**		have do	yes, i need a 4 star hotel that also
	free parking	parking	includes free wifi.
***Cluster* 12**			can you give me the address, hotel type, and phone number for either of
	looking for	look for	the hotels?
***Cluster* 13**	to		oh, of course. sorry. i'm leaving out of
	cambridge 1	look be for 3	leicester and going into cambridge.
***Cluster* 14**			thanks. i will also need a taxi from the hotel to the restaurant. will you handle
	taxi to	need book	this?
***Cluster* 16**			no, that's all i needed today. thanks for
	thank you 2	thank you 1	your help, it's much appreciated.
***Cluster* 17**	thank you 3	thank you 2	that is all i need.
***Cluster* 18**	phone		yes, that will work great. can i get their
	number	be what 2	phone number please?
***Cluster* 19**		be there	several restaurants fit your criteria. do
	in the	restaurant	you prefer a specific area?
***Cluster* 20**			the postcode is cb58bs. is there
	postcode is	be and be	anything else i can help you with?
***Cluster* 21**	have great	have day	you're welcome - have a great day!
***Cluster* 22**			
***Cluster* 23**	will you	be what	sure! what is your departure site?
***Cluster* 24**			tr9025 arrives at liverpool st at 13:27.
	you like	be minute	would you like to book a ticket?

***Cluster* 25**			yes, there are 11 museums. i'd recommend broughton house gallery. it's at 98 king street cb11ln and has free entrance. would that work for you?
	in the 1	be in	
***Cluster* 26**			i have 133 trains matching your request. is there a specific day and time you would like to travel?
	you like 1	be there train	
***Cluster* 27**	anything else	be there anything	no problem. is there anything else i can help you with today?
***Cluster* 28**			i have two hotels that match the description of 3 star ratings. are you looking for a hotel in the centre or south area?
	in the 2	be hotel	
***Cluster* 29**	cambridge towninfo	thank you for have day	thank you for using the cambridge towneinfo help desk!
***Cluster* 30**	reference number	be reserve table will	i've booked it for you! the reference is lqjwqd08 .
***Cluster* 31**			booking completed! booked car type : yellow bmw contact number : 07546378894
	number is	be	
***Cluster* 32**	you like 2	have do range	no. they all have 4 stars.
***Cluster* 33**		be there attraction	okay. and did you want it in the west or the south?
	in the 3	information	
***Cluster* 34**	reference number 1	be be payable be	the price per ticket is 10.10 pounds.