Proposta recebida em 30 de Outubro 2018 e aceite para publicação em 12 de Julho 2019.

Reconhecimento de Actos de Diálogo Hierárquicos e Multi-Etiqueta em Dados em Espanhol

Hierarchical Multi-Label Dialog Act Recognition on Spanish Data

Eugénio Ribeiro
L^{2}F – Spoken Language Systems Laboratory – INESC-ID Lisboa
Instituto Superior Técnico, Universidade de Lisboa, Portugal
eugenio.ribeiro@inesc-id.pt

Ricardo Ribeiro
L^{2}F – Spoken Language Systems Laboratory – INESC-ID Lisboa
ISCTE-IUL – Instituto Universitário de Lisboa, Portugal
ricardo.ribeiro@inesc-id.pt

David Martins de Matos
L^{2}F – Spoken Language Systems Laboratory – INESC-ID Lisboa
Instituto Superior Técnico, Universidade de Lisboa, Portugal
david.matos@inesc-id.pt

Resumo

Os actos de diálogo revelam a intenção por trás das palavras pronunciadas. Por isso, o seu reconhecimento automático é importante para um sistema de diálogo que tenta entender o seu interlocutor. O estudo apresentado neste artigo aborda essa tarefa no corpus DIHANA, cujo esquema de anotação de actos de diálogo em três níveis coloca problemas que não foram explorados em estudos recentes. Além do problema hierárquico, os dois níveis inferiores colocam problemas de classificação multi-etiqueta. Além disso, cada nível da hierarquia refere-se a um aspecto diferente relativo à intenção do orador, tanto em termos da estrutura do diálogo, como da tarefa. Por outro lado, uma vez que os diálogos são em espanhol, este corpus permite-nos avaliar se as melhores abordagens para dados em inglês generalizam para uma língua diferente. Mais especificamente, comparamos o desempenho de diferentes abordagens de representação de segmentos, com foco tanto em sequências como em padrões de palavras, e avaliamos a importância do histórico do diálogo e das relações entre os múltiplos níveis da hierarquia. No que diz respeito ao problema de classificação de etiqueta única colocado pelo nível superior, mostramos que as conclusões obtidas a partir de dados em inglês se mantêm em dados em espanhol. Para além disso, mostramos que as abordagens podem ser adaptadas para cenários multi-etiqueta. Por fim, combinando hierarquicamente os melhores classificadores para cada nível, obtemos os melhores resultados reportados para este corpus.

Palavras chave

reconhecimento de actos de diálogo, classificação hierárquica, classificação multi-etiqueta

Abstract

Dialog acts reveal the intention behind the uttered words. Thus, their automatic recognition is important for a dialog system trying to understand its conversational partner. The study presented in this article approaches that task on the DIHANA corpus, whose three-level dialog act annotation scheme poses problems which have not been explored in recent studies. In addition to the hierarchical problem, the two lower levels pose multi-label classification problems. Furthermore, each level in the hierarchy refers to a different aspect concerning the intention of the speaker both in terms of the structure of the dialog and the task. Also, since its dialogs are in Spanish, it allows us to assess whether the state-of-the-art approaches on English data generalize to a different language. More specifically, we compare the performance of different segment representation approaches focusing on both sequences and patterns of words and assess the importance of the dialog history and the relations between the multiple levels of the hierarchy. Concerning the single-label classification problem posed by the top level, we show that the conclusions drawn on English data also hold on Spanish data. Furthermore, we show that the approaches can be adapted to multi-label scenarios. Finally, by hierarchically combining the best classifiers for each level, we achieve the best results reported for this corpus.

Keywords

dialog act recognition, hierarchical classification, multi-label classification

1 Introdução

Para um sistema de diálogo é relevante identificar a intenção por trás das palavras dos seus interlocutores, uma vez que esta fornece uma pista importante sobre a informação contida num segmento e como este deve ser interpretado. Segundo Searle (1969), essa intenção é revelada pelos actos de diálogo, que são as unidades mínimas de comunicação linguística. Consequentemente, o reconhecimento automático de actos de diálogo é uma tarefa importante no contexto da compreensão de língua natural, que tem sido amplamente explorada ao longo dos anos em múltiplos corpora com diferentes características. Recentemente, a maioria dos estudos tem-se focado em dados em inglês e, mais especificamente, no Switchboard Dialog Act Corpus (SwDA) (Jurafsky et al., 1997), uma vez que este é o maior corpus anotado com actos de diálogo e o seu conjunto de etiquetas é independente da tarefa e do domínio. No entanto, existem outros corpora e esquemas de anotação que colocam problemas no contexto do reconhecimento de actos de diálogo que não são cobertos pelo corpus SwDA e as suas anotações no formato SWBD-DAMSL. Tendo isso em conta, neste artigo exploramos o corpus DIHANA (Benedí et al., 2006), que contém interações em espanhol entre humanos e um sistema de diálogo simulado usando o método do Feiticeiro de Oz (WoZ). No contexto do reconhecimento de actos de diálogo, este corpus diferencia-se dos restantes pelo seu esquema de anotação em três níveis, no qual o nível superior se refere ao acto de diálogo genérico e independente da tarefa e os restantes o complementam com informação específica da tarefa. Para além disso, cada segmento tem apenas uma etiqueta de nível superior, mas pode não ter nenhuma ou ter várias etiquetas nos restantes níveis. Tendo em conta estas características, o corpus DIHANA permite-nos abordar o reconhecimento de actos de diálogo como um problema de classificação hierárquica e multi-etiqueta.

Similarmente ao que acontece com outras tarefas de classificação de texto, tais como categorização de notícias e análise de sentimento (Kim, 2014; Conneau et al., 2017), a maioria das abordagens recentes ao reconhecimento de actos de diálogo baseiam-se em Redes Neuronais Profundas (DNNs). Uma visão geral sobre essas abordagens é fornecida na Secção 2.2. No entanto, em geral, estas usam uma abordagem baseada em Redes Neuronais Recorrentes (RNNs) ou Redes Neuronais Convolucionais (CNNs) para gerar uma representação do segmento a partir da representação das suas palavras na forma de *embeddings*. Em seguida, a informação presente nessa representação é usada para obter a classificação do segmento. A distinção entre as abordagens baseadas em RNNs e CNNs é relevante, uma vez que estas são capazes de capturar diferentes tipos de informação. No caso das primeiras, o foco é em identificar sequências de palavras relevantes, incluindo dependências de longo alcance. Por outro lado, as últimas focam-se na identificação de padrões de palavras relevantes, observando janelas de contexto em torno de cada palavra. Para além disso, as abordagens com desempenho mais alto na tarefa não consideram cada segmento por si só, mas sim em conjunto com informação de contexto extraída dos segmentos circundantes e sobre os oradores.

Tendo em conta as características do corpus DIHANA e as melhores abordagens para o reconhecimento automático de actos de diálogo independentes do domínio e com apenas uma etiqueta, neste artigo exploramos diferentes aspectos relacionados com a tarefa. Em primeiro lugar, avaliamos se essas abordagens têm um desempenho semelhante numa língua diferente do inglês, utilizando-as para prever as etiquetas independentes do domínio do nível superior. Em seguida, exploramos a sua aplicabilidade nos cenários de classificação multi-etiqueta colocados pelos restantes níveis. Para além disso, uma vez que esses níveis se referem a diferentes aspectos específicos da tarefa, também avaliamos como a informação de contexto extraída dos segmentos anteriores influencia a capacidade de prever cada um desses aspectos. Da mesma forma, avaliamos como essa habilidade é influenciada por informação relativa aos níveis superiores da hierarquia. Por fim, exploramos a combinação hierárquica das melhores abordagens para cada nível e comparamos o seu desempenho com o da abordagem plana que foi utilizada em estudos anteriores sobre o mesmo corpus.

No resto do artigo, começamos por fornecer uma visão geral sobre o trabalho relacionado na Secção 2. Nesse sentido, começamos por fornecer uma visão geral sobre corpora para o reconhecimento de actos de diálogo na Secção 2.1. Em seguida, discutimos as melhores abordagens para o reconhecimento de actos de diálogo na Secção 2.2. Adicionalmente, resumimos estudos anteriores sobre o reconhecimento de actos de diálogo em dados em espanhol na Secção 2.3. Após essa discussão, na Secção 3, descrevemos a nossa configuração experimental. Começamos por descrever o corpus DIHANA e as suas anotações de actos de diálogo na Secção 3.1. A Secção 3.2 apresenta a arquitectura genérica das redes utilizadas nas

nossas experiências e descreve o que muda entre cada uma delas. Por fim, a Secção 3.3 descreve os procedimentos de treino e avaliação de acordo com o nível da hierarquia em foco. Os resultados alcançados pelas nossas experiências em cada um desses níveis, assim como na sua combinação, são apresentados e discutidos na Secção 4. Por fim, a Secção 5 apresenta as conclusões mais importantes que podem ser tiradas das experiências descritas neste artigo e fornece indicadores para trabalho futuro.

2 Trabalho Relacionado

Como mencionado anteriormente, o reconhecimento automático de actos de diálogo é uma tarefa que tem sido amplamente explorada ao longo dos anos em múltiplos corpora com diferentes características e usando uma grande variedade de técnicas de aprendizagem clássicas, desde Modelos de Markov Ocultos (HMMs) (Stolcke et al., 2000) até Máquinas de Vectores de Suporte (SVMs) (Gambäck et al., 2011). O artigo de Král & Cerisara (2010) fornece uma visão geral sobre a maioria dessas abordagens. No entanto, recentemente, a maioria das abordagens baseia-se em diferentes arquitecturas de DNNs. Abaixo, apresentamos um sumário dessas abordagens. Para além disso, uma vez que o nosso estudo se foca no corpus DIHANA (Benedí et al., 2006), temos também uma subsecção dedicada a abordagens aplicadas no reconhecimento de actos de diálogo em dados em espanhol. No entanto, antes de discutirmos abordagens, fornecemos uma visão geral sobre corpora existentes para o reconhecimento de actos de diálogo.

2.1 Corpora para Reconhecimento de Actos de Diálogo

Vários corpora foram anotados com actos de diálogo. A Tabela 1 apresenta um conjunto não exaustivo desses corpora e das suas características. Podemos ver que múltiplos domínios, linguagens e tipos de interação são cobertos, o que permite a avaliação das capacidades de generalização das abordagens de reconhecimento de actos de diálogo para múltiplos cenários. No entanto, por outro lado, os conjuntos de etiquetas utilizados não são padronizados entre corpora. De facto, existem até conjuntos distintos de etiquetas para o mesmo corpus. Isso significa que esses conjuntos foram desenvolvidos com diferentes objectivos e têm diferentes hierarquias e níveis de abstracção, o que dificulta a realização de experiências de generalização entre corpora. Isto é particularmente problemático quando os conjuntos de etiquetas usados são dependentes do domínio, uma vez que não podem ser aplicados a corpora noutros domínios.

Em relação a conjuntos alternativos de etiquetas para o mesmo corpus, enquanto os dos corpora SwDA, ICSI Meeting Recorder Dialog Act Corpus (MRDA) e CallHome Spanish (CHS) são apenas versões comprimidas dos conjuntos originais, os dois conjuntos de etiquetas usados para anotar o corpus VERBMOBIL são disjuntos. Para além disso, o primeiro inclui etiquetas dependentes do domínio (Jekat et al., 1995), enquanto o segundo é completamente independente do domínio (Alexandersson et al., 1998).

Múltiplos corpora têm conjuntos de etiquetas complementares que se referem a diferentes aspectos. Por exemplo, os corpora MRDA, DIHANA e NESPOLE têm um conjunto de etiquetas genéricas que podem ser especializadas usando etiquetas de outros conjuntos. No entanto, enquanto no primeiro caso as etiquetas especializadas ainda são independentes do domínio, nos dois restantes as etiquetas genéricas são complementadas com informação específica do domínio a diferentes níveis. No corpus DIME, os dois conjuntos de etiquetas referem-se a diferentes aspectos do diálogo, nomeadamente, definição de obrigações e estabelecimento de uma base comum. Por último, o corpus LEGO tem conjuntos de etiquetas independentes para os segmentos do utilizador e do sistema.

Numa tentativa de padronizar a anotação de actos de diálogo e, consequentemente, estabelecer uma base para estudos mais comparáveis na área, Bunt et al. (2012) definiram a norma ISO 24617-2. De acordo com esta norma, as anotações de actos de diálogo devem ser realizadas em segmentos funcionais, em vez de em turnos ou frases (Carroll & Tanenhaus, 1978). Para além disso, a anotação de cada segmento não consiste apenas numa etiqueta, mas sim numa estrutura complexa contendo informação sobre os participantes, relações com outros segmentos funcionais, a dimensão semântica do acto de diálogo, a sua função comunicativa e qualificadores opcionais sobre certeza, condicionalidade, parcialidade e sentimento. No entanto, anotar todos estes aspectos é um processo exaustivo e, consequentemente, a quantidade de dados anotados de acordo com a norma é ainda reduzida e, em muitos casos, nem todos os aspectos são considerados (Petukhova et al., 2014; Bunt et al., 2016; Ribeiro et al., 2016).

Como mencionado anteriormente, os estudos mais recentes sobre o reconhecimento automático

Corpus	Interacção	Domínio	Língua	Segmentos	Etiquetas	DD
SwDA (Jurafsky et al., 1997)	Humanos	Aberto	Inglês	220k	41 - 44	N
MRDA (Shriberg et al., 2004)	Humanos	Reuniões	Inglês	106k	5 / 11 + 39	N
AMI (Carletta et al., 2005)	Humanos	Reuniões	Inglês	102k	15	N
VERBMOBIL (Kay et al., 1992)	Humanos	Horários	Múltiplas	59k	42 / 33	M
CHS (Levin et al., 1998)	Humanos	Aberto	Espanhol	45k	10 / 37	N
DSTC4 (Kim et al., 2016)	Humanos	Viagens	Inglês	31k	89	S
MapTask (Anderson et al., 1991)	Humanos	Mapas	Inglês	27k	12	N
DIHANA (Benedí et al., 2006)	WoZ	Comboios	Espanhol	23k	11 + 10 + 13	M
LEGO (Schmitt et al., 2012)	Máquina	Autocarros	Inglês	14k	22 + 28	S
NESPOLE (Costantini et al., 2002)	Humanos	Viagens	Múltiplas	8k	67 + 91	M
DIME (Villaseñor et al., 2001)	WoZ	Cozinhas	Espanhol	5k	15 + 15	M

Tabela 1: Corpora anotados com actos de diálogo, ordenados por número aproximado de segmentos. A coluna referente à interacção diz se os diálogos são entre humanos ou existe um sistema de diálogo envolvido. No último caso, são distinguidos os cenários que usam o método WoZ daqueles que envolvem interacção real com uma máquina. Na coluna referente ao número de etiquetas, os símbolos / e - referem-se a conjuntos alternativos de etiquetas, enquanto o símbolo + se refere a diferentes níveis de anotação. A última coluna diz se o conjunto de etiquetas é dependente do domínio (S), independente do mesmo (N), ou se existem etiquetas de ambos os tipos (M).

de actos de diálogo utilizam diferentes arquitecturas de DNN. Tais abordagens requerem grandes quantidades de dados para serem treinadas. Consequentemente, a identificação automática de actos de diálogo como definidos pela norma ISO só foi abordada num conjunto reduzido de estudos (Ribeiro et al., 2015; Mezza et al., 2018). Por outro lado, o corpus SwDA é o mais explorado para a tarefa, uma vez que é aquele que possui o maior número de segmentos anotados, os seus diálogos cobrem múltiplos domínios e seu conjunto de etiquetas é independente do domínio. Por isso, é esperado que as conclusões tiradas de experiências sobre este corpus generalizem bem para outros cenários.

2.2 Estado da Arte em Reconhecimento de Actos de Diálogo

As abordagens com melhor desempenho na tarefa de reconhecimento de actos de diálogo são baseadas em DNNs. Por isso, nesta secção, focamo-nos em estudos que usam esse tipo de abordagem. Pelo que sabemos, o primeiro desses estudos foi o de Kalchbrenner & Blunsom (2013). O método descrito utiliza uma abordagem baseada em CNNs para gerar a representação de um segmento a partir da representação das suas palavras na forma de *embeddings* inicializados aleatoriamente. Em seguida, é usado um modelo de discurso baseado em RNNs que combina a sequência de representações de segmentos com informação sobre os oradores e produz a sequência de actos de diálogo correspondente. Ao limitar o modelo de discurso para considerar informação de apenas dois segmentos anteriores, esta abordagem alcançou uma taxa de acerto de 73,9 % no corpus SwDA.

Lee & Dernoncourt (2016) compararam o desempenho de uma unidade recorrente de Longa Memória de Curto Prazo (LSTM) com o de uma CNN para gerar representações de segmentos a partir da representação das suas palavras na forma de *embeddings* pré-treinados. Para identificar os actos de diálogo correspondentes, as representações de segmentos são passadas por uma rede totalmente ligada com duas camadas, na qual a primeira normaliza as representações e a segunda seleciona a classe com maior probabilidade. Nas experiências realizadas, a abordagem baseada em CNNs levou consistentemente a resultados semelhantes ou melhores do que aqueles da abordagem baseada em LSTM. A arquitectura foi também adaptada para fornecer informações de contexto a dois níveis e de até dois segmentos anteriores. O primeiro nível refere-se à concatenação das representações dos segmentos precedentes com a do segmento atual antes de o fornecer à rede totalmente ligada. O segundo refere-se à concatenação das representações normalizadas antes de serem fornecidas à camada de saída. Esta abordagem alcançou uma taxa de acerto de 65,8% no corpus Dialog State Tracking Challenge 4 (DSTC4), 84,6% no corpus MRDA com cinco classes (Ang et al., 2005) e 71,4% no corpus SwDA. No entanto, a influência da informação de contexto variou entre corpora.

Ji et al. (2016) exploraram a combinação de aspectos positivos de redes neuronais e modelos gráficos probabilísticos. Eles usaram um Modelo de Língua com Relações de Discurso (DRLM)

que combina um Modelo de Língua Baseado em RNNs (RNNLM) (Mikolov et al., 2010), para modelar a sequência de palavras no diálogo, com um modelo de variável latente sobre a estrutura do discurso, para modelar relações entre segmentos adjacentes que, neste contexto, representam os actos de diálogo. Desta forma, o modelo pode prever palavras usando representações vectoriais treinadas de forma discriminativa enquanto mantém uma representação probabilística de um elemento linguístico alvo, como o acto de diálogo. Para funcionar como um classificador de actos de diálogo, o modelo foi treinado para maximizar a probabilidade condicional de uma sequência de actos de diálogo dada uma sequência de segmentos, alcançando uma taxa de acerto de 77,0% no corpus SwDA.

Tran et al. (2017b) usaram uma RNN hierárquica com um mecanismo de atenção para prever as classificações de actos de diálogo de um diálogo inteiro. O modelo é hierárquico, uma vez que inclui uma RNN ao nível do segmento, para gerar a sua representação a partir das suas palavras, e outra para gerar a sequência de etiquetas de acto de diálogo a partir da sequência de representações de segmento. O mecanismo de atenção está entre os dois, uma vez que usa informações da RNN ao nível do diálogo para identificar as palavras mais importantes no segmento actual e filtrar a sua representação. Usando esta abordagem eles alcançaram uma taxa de acerto de 74,5% no corpus SwDA e 63,3% no corpus HCRC Map Task Corpus (MapTask). Mais tarde, o desempenho no corpus SwDA foi melhorado para 75,6% usando um método baseado na propagação de informação de incerteza sobre as previsões anteriores (Tran et al., 2017c). Para além disso, utilizando mecanismos de atenção aplicados às células das camadas recorrentes no contexto de um modelo generativo, alcançaram uma taxa de acerto de 74,2% no corpus SwDA e 65,94% no corpus MapTask (Tran et al., 2017a).

Os estudos referidos anteriormente exploraram a utilização de uma única camada recorrente ou convolucional para gerar a representação do segmento a partir das suas palavras. No entanto, as abordagens com melhor desempenho na tarefa utilizam múltiplas dessas camadas. Por um lado, Khanpour et al. (2016) alcançaram os seus melhores resultados usando uma representação de segmento gerada pela concatenação das saídas de uma pilha de 10 unidades LSTM. Deste modo, o modelo é capaz de capturar relações de longa distância entre palavras. Por outro lado, Liu et al. (2017) geraram a representação do segmento combinando as saídas de três CNNs paralelas com diferentes tamanhos de janela de contexto, para capturar diferentes padrões funcionais. Em ambos os casos, representações das palavras na forma de *embeddings* pré-treinados foram usadas como entrada para a rede. Em geral, a partir dos resultados reportados, não é possível afirmar qual é a abordagem de representação de segmento com melhor desempenho, uma vez que a avaliação foi realizada em diferentes subconjuntos do corpus SwDA. Ainda assim, Khanpour et al. (2016) atingiram uma taxa de acerto de 73,9% no conjunto de validação e 80,1% no conjunto de teste, enquanto Liu et al. (2017) atingiram taxas de acerto de 74,5% e 76,9% nos dois conjuntos utilizados para avaliar as suas experiências. Para além disso, Khanpour et al. (2016) atingiram uma taxa de acerto de 86,8% no corpus MRDA.

Liu et al. (2017) exploraram também o uso de informação de contexto sobre mudança de orador e extraída dos segmentos circundantes. Para fornecer informação sobre a mudança de orador limitaram-se a adicionar um valor binário à representação do segmento, que indica se o orador mudou em relação ao segmento anterior. Já em relação a informação dos segmentos circundantes, exploraram o uso de modelos de discurso, assim como de abordagens que concatenam a informação de contexto directamente na representação do segmento. Os modelos de discurso tornam o modelo hierárquico, gerando uma sequência de classificações de actos de diálogo a partir da sequência de representações de segmento. Assim, ao prever a classificação de um segmento, aqueles que o circundam também são levados em conta. No entanto, quando o modelo de discurso é baseado numa CNN ou numa unidade LSTM bidireccional, ele considera informação de segmentos futuros, que não estão disponíveis para um sistema de diálogo. Ainda assim, mesmo tendo em conta informação futura, as abordagens baseadas em modelos de discurso tiveram pior desempenho do que aquelas que concatenaram a informação de contexto directamente na representação do segmento. Nesse aspecto, fornecer essa informação na forma das classificações de acto de diálogo dos segmentos circundantes levou a melhores resultados do que utilizar as suas palavras, mesmo quando essas classificações foram obtidas automaticamente. Esta conclusão está alinhada com o que tínhamos mostrado no nosso estudo anterior, utilizando SVMs (Ribeiro et al., 2015). Para além disso, ambos os estudos demonstraram que, como esperado, o primeiro segmento anterior é o mais importante e que a influência diminui com a distância. Usando as etiquetas de referência de

três segmentos anteriores, os resultados nos dois conjuntos usados para avaliar a abordagem melhoraram para 79,6% e 81,8%, respectivamente.

É importante fazer algumas observações sobre tokenização e representação de *tokens*. Em todos os estudos descritos anteriormente, a tokenização foi realizada no nível da palavra. Para além disso, com excepção do primeiro estudo (Kalchbrenner & Blunsom, 2013), em que foram utilizadas representações na forma de *embeddings* inicializados aleatoriamente, e dos realizados por Tran et al. (2017a,b,c), para os quais a abordagem de representação não é revelada nos artigos, a representação dessas palavras é feita na forma de *embeddings* pré-treinados. Khanpour et al. (2016) compararam a performance utilizando *embeddings* treinados usando os métodos Word2Vec (Mikolov et al., 2013) e Vectores Globais para Representação de Palavras (GloVe) (Pennington et al., 2014) em múltiplos corpora. Embora ambas as abordagens capturem informação relativa a palavras que aparecem juntas frequentemente, os melhores resultados foram alcançados usando a abordagem Word2Vec. Em termos de dimensionalidade, Khanpour et al. (2016) alcançou os melhores resultados ao usar *embeddings* com 150 dimensões. No entanto, outros estudos (Lee & Dernoncourt, 2016; Liu et al., 2017) usam *embeddings* com 200 dimensões, sendo que este não foi um dos valores comparados para a dimensionalidade.

As abordagens descritas em todos os estudos referidos anteriormente realizam tokenização ao nível da palavra. No entanto, num estudo recente (Ribeiro et al., 2018), mostrámos que também existem pistas importantes para a detecção de intenção a um nível sub-palavra, que só podem ser capturadas quando se usa uma tokenização mais granular, como, por exemplo, ao nível do carácter (Ribeiro et al., 2018). As pistas a esse nível referem-se principalmente a aspectos relativos à morfologia das palavras, tais como lemas e afixos. Para capturar essa informação, nós adaptámos a abordagem de representação de segmento baseada em CNNs descrita por Liu et al. (2017) para usar caracteres em vez de palavras. Dessa forma, pudemos explorar janelas de contexto de diferentes tamanhos para capturar diferentes aspectos morfológicos. Neste sentido, os nossos melhores resultados foram alcançados quando utilizámos três CNNs paralelas com janelas de tamanho três, cinco e sete, que são capazes de capturar afixos, lemas e relações entre palavras, respectivamente. Usando essa abordagem, obtivemos taxas de acerto de 76,88% e 73,22% nos conjuntos de validação e teste do corpus SwDA, respectivamente. Estes resultados são semelhantes aos da abordagem ao nível da palavra. No entanto, a combinação dos dois níveis melhorou os resultados para 78,0% e 74,0%, respectivamente, o que mostra que estes capturam informação complementar. Por fim, ao incluir informação de contexto de três segmentos anteriores, melhorámos os resultados para 82,0% no conjunto de validação e 79,0% no conjunto de teste.

2.3 Reconhecimento de Actos de Diálogo em Dados em Espanhol

A investigação sobre o reconhecimento de actos de diálogo em dados em espanhol tem sido realizada principalmente em dois corpora — DIHANA e CHS. Em ambos, os diálogos são telefónicos e espontâneos. No entanto, tal como mostrado na Tabela 1, enquanto os diálogos do primeiro são entre humanos e um sistema de diálogo, os do segundo são entre humanos. Para além disso, enquanto o corpus CHS está anotado com etiquetas independentes do domínio e da tarefa, o corpus DIHANA está anotado segundo um esquema hierárquico com três níveis, em que o primeiro se refere ao acto de diálogo genérico e independente do domínio e os restantes o complementam com informação específica da tarefa. Existe também uma série de trabalhos sobre reconhecimento de actos de diálogo no corpus DIME (Coria & Pineda, 2005, 2006, 2009). No entanto, estes focam-se em usar informação prosódica para prever subconjuntos específicos dos actos de diálogo relacionados com obrigações e estabelecimento de uma base comum com que o corpus está anotado. Uma vez que o nosso trabalho se foca no reconhecimento de actos de diálogo a partir de informação textual, apenas vamos descrever mais detalhadamente os estudos sobre os dois primeiros corpora.

Os primeiros estudos em reconhecimento de actos de diálogo no corpus DIHANA usaram HMMs, quer baseados em informação prosódica (Tamarit & Martínez-Hinarejos, 2008) — energia e frequência fundamental —, quer textual (Martínez-Hinarejos et al., 2008) — n-gramas de palavras. O primeiro atingiu uma taxa de acerto de 60,70% no primeiro nível, enquanto o segundo alcançou 93,40% na combinação dos dois primeiros níveis e 89,70% na combinação de todos os níveis. Este segundo estudo e um outro mais recente (Martínez-Hinarejos et al., 2015) também exploraram o reconhecimento de actos de diálogo em diálogos não segmentados à priori, usando transdutores de n-gramas. No entanto, nesses casos, o foco foi no processo de segmentação e as

abordagens de classificação actos de diálogo não diferiram das anteriores. Por fim, os melhores resultados nos diálogos segmentados manualmente foram obtidos usando uma abordagem baseada em SVMs aplicados a n-gramas de palavras, informação sobre a presença de palavras de pergunta e pontuação, e informação de contexto de três segmentos anteriores (Gambäck et al., 2011). Para além disso, foi também aplicada uma abordagem de aprendizagem activa para reduzir a quantidade de dados necessários para o treino, alcançando uma taxa de acerto de 94,08% na combinação dos dois primeiros níveis e 90,97% na combinação de todos os níveis.

Tal como no corpus DIHANA, os primeiros estudos em reconhecimento de actos de diálogo no corpus CHS também usaram HMMs com diferentes tipos de n-grama (Levin et al., 1999; Ries, 1999). O segundo estudo referido melhorou os resultados ao combinar os HMMs com redes neuronais aplicadas a unigramas etiquetas de Parte do Discurso (POS), alcançando uma taxa de acerto de 76.1%. A tarefa também foi abordada usando Análise de Semântica Latente (LSA) em três estudos diferentes (Serafin et al., 2003; Serafin & Di Eugenio, 2004; Di Eugenio et al., 2010). O primeiro usou não só LSA básica, como também múltiplas adaptações baseadas em aglomeração e na incorporação de informação relativa aos actos de diálogo anteriores. No entanto, não foi observada uma melhoria significativa em relação à LSA básica, que alcançou uma taxa de acerto de 65,36% no conjunto com 37 etiquetas 68,91% na sua versão comprimida, com 10 etiquetas. Por outro lado, os restantes estudos exploraram o uso de informação sobre múltiplas características sintáticas e relacionadas com o diálogo, atingindo resultados superiores aos obtidos usando LSA básica, até um máximo de 77,74% e 81,27%, respectivamente. No último estudo, esses resultados foram ainda melhorados para 80,34% e 82,88%, através da aplicação de uma abordagem de aprendizagem baseada em instâncias, mais especificamente, k-Vizinhos Mais Próximos (k-NN), aos espaços semânticos obtidos através da LSA. No entanto, em ambos os casos, as melhorias foram alcançadas utilizando informação relativa ao objectivo do diálogo, ou seja, a intenção genérica por trás de todo o diálogo, e sobre se o orador está a tomar a iniciativa, ou apenas a responder ou a acompanhar o outro orador. Embora em geral o objectivo do diálogo seja conhecido, existem alguns casos em que um sistema de diálogo não tem essa informação. Para além disso, identificar se um orador está a tomar a iniciativa, a responder ou a acompanhar o outro orador pode ser visto como uma simplificação da tarefa de reconhecimento de actos de diálogo. Logo, não é justo usar essa informação caso esta não seja obtida automaticamente também. Por fim, o corpus CHS também foi explorado em experiências de adaptação a diferentes domínios para classificação de actos de diálogo usando um conjunto reduzido de classes (Margolis et al., 2010).

3 Configuração Experimental

Queremos avaliar se as abordagens com melhor desempenho descritas na secção anterior têm um desempenho semelhante numa língua diferente do inglês. Para além disso, queremos explorar a sua aplicabilidade nos cenários de classificação multi-etiqueta colocados pelos dois níveis inferiores das anotações de actos do diálogo do corpus DIHANA. Como esses níveis se referem a diferentes aspectos específicos da tarefa, também avaliamos como a informação de contexto extraída dos segmentos anteriores influencia a capacidade de prever cada um desses aspectos. Da mesma forma, avaliamos como essa capacidade é influenciada por informação dos níveis acima na hierarquia. Por fim, queremos avaliar se a combinação hierárquica das melhores abordagens para cada nível é capaz de superar a abordagem plana utilizada em estudos anteriores sobre o mesmo corpus.

Nesta secção descrevemos a nossa configuração experimental, começando com uma descrição do corpus DIHANA e das suas anotações de acto de diálogo. Em seguida, apresentamos a arquitectura genérica usada nas nossas experiências e explicamos como ela muda de acordo com o aspecto e as características do nível em foco, especialmente considerando as diferenças entre os cenários de classificação com etiqueta única e os de classificação multi-etiqueta. Por fim, descrevemos as nossas abordagens de treino e avaliação, incluindo as diferenças nas métricas usadas para problemas com etiqueta única e multi-etiqueta.

3.1 Corpus

O corpus DIHANA (Benedí et al., 2006) consiste em 900 diálogos telefónicos entre 225 humanos e um sistema de diálogo que fornece informação sobre comboios, simulado usando o método WoZ. Existem 6.280 turnos de utilizadores e 9.133 turnos do sistema, com um vocabulário de 823 palavras e um total de 48.243 palavras. Os turnos foram transcritos, segmentados e anotados manualmente com actos de diálogo (Alcácer et al.,

SISTEMA: Bienvenido al servicio de informacion de trenes ¿En qué puedo ayudarle?
(Bem-vindo ao serviço de informação de comboios. Como posso ajudá-lo?)
Nível 1: *Apertura*, Nível 2: *Nula*, Nível 3: *Nula*

UTILIZADOR: Quería ir a La Coruña *(Queria ir à Corunha)*
Nível 1: *Pregunta*, Nível 2: *Hora Salida*, Nível 3: *Destino*

UTILIZADOR: y quería saber los horarios que hay de los trenes para el lunes.
(e gostaria de saber o horário dos comboios à segunda-feira.)
Nível 1: *Pregunta*, Nível 2: *Hora Salida*, Nível 3: *Día*

SISTEMA: ¿Quiere viajar desde desde La Coruña?
(Quer viajar a partir da Corunha?)
Nível 1: *Confirmación*, Nível 2: *Origen*, Nível 3: *Origen*

UTILIZADOR: No. *(Não.)*
Nível 1: *Negación*, Nível 2: *Origen*, Nível 3: *Nula*

UTILIZADOR: Desde Bilbao. *(A partir de Bilbau.)*
Nível 1: *Respuesta*, Nível 2: *Hora Salida*, Nível 3: *Origen*

SISTEMA: Quiere horarios de trenes a La Coruña
(Quer horários de comboios para a Corunha)
Nível 1: *Confirmación*, Nível 2: *Hora Salida*, Nível 3: *Destino*

SISTEMA: ¿desde Bilbao? *(a partir de Bilbau?)*
Nível 1: *Confirmación*, Nível 2: *Origen*, Nível 3: *Origen*

UTILIZADOR: Sí. *(Sim.)*
Nível 1: *Afirmación*, Nível 2: *Origen*, Nível 3: *Nula*

SISTEMA: Le consulto horarios de trenes con salida en Bilbao destino La Coruña saliendo el lunes, día 16 de febrero de 2004.
(Estou à procura de comboios de Bilbau para a Corunha na segunda-feira, 16 de Fevereiro de 2004.)
Nível 1: *Confirmación*, Nível 2: *Hora Salida*, Nível 3: *Destino, Día, Origen*

SISTEMA: Un momento por favor. *(Um momento por favor.)*
Nível 1: *Espera*, Nível 2: *Nula*, Nível 3: *Nula*

SISTEMA: El único tren que realiza el trayecto es un Diurno que sale a las 9 y 25 de la mañana.
(O único comboio que faz esse percurso é um Diurno que parte às 9 e 25 da manhã.)
Nível 1: *Respuesta*, Nível 2: *Hora Salida*, Nível 3: *Hora Salida,Número Trenes,Tipo Tren*

SISTEMA: ¿Desea algo más? *(Precisa de mais alguma coisa?)*
Nível 1: *Nueva Consulta*, Nível 2: *Nula*, Nível 3: *Nula*

UTILIZADOR: Sí, ¿me puede decir el precio del billete?
(Sim. Pode dizer-me o preço do bilhete?)
Nível 1: *Pregunta*, Nível 2: *Precio*, Nível 3: *Nula*

SISTEMA: El precio de ese tren en clase turista es 36 euros el billete de ida. Recuerde que el billete de ida y vuelta siempre supone un ahorro del 20 por ciento.
(O preço do bilhete de ida nesse comboio é de 36 euros em classe económica. Lembre-se que bilhetes de ida e volta têm um desconto de 20%.)
Nível 1: *Respuesta*, Nível 2: *Precio*, Nível 3: *Clase Billete, Precio*

SISTEMA: ¿Desea algo más? *(Precisa de mais alguma coisa?)*
Nível 1: *Nueva Consulta*, Nível 2: *Nula*, Nível 3: *Nula*

UTILIZADOR: No. Creo que no. *(Não. Penso que não.)*
Nível 1: *Encerramento*, Nível 2: *Nula*, Nível 3: *Nula*

SISTEMA: Gracias por utilizar este servicio. Feliz viaje.
(Obrigado por usar este serviço. Boa viagem.)
Nível 1: *Encerramento*, Nível 2: *Nula*, Nível 3: *Nula*

Figura 1: Um diálogo do corpus DIHANA. Cada segmento é seguido pela sua tradução em português, entre parêntesis, e pelas suas etiquetas de actos de diálogo em cada um dos três níveis.

2005). O número total de segmentos anotados é 23.547, 9.715 dos quais são de utilizadores e 13.832 do sistema. A Figura 1 mostra um exemplo de um diálogo anotado.

As anotações de actos de diálogo são decompostas hierarquicamente em três níveis (Martínez-Hinarejos et al., 2002). Enquanto o primeiro nível representa a intenção genérica do segmento, independente de detalhes relativos ao domínio e à tarefa, os restantes representam informação específica da tarefa. O primeiro nível tem 11 etiquetas, distribuídas de acordo com a Tabela 2. Nessa tabela podemos ver que duas das etiquetas são exclusivas a segmentos de utilizadores — *Aceitação* e *Rejeição* — e quatro a segmentos do sistema — *Abertura*, *Espera*, *Nova Consulta* e *Confirmação*. Para além disso, a etiqueta mais comum, *Pergunta*, cobre 27% dos segmentos.

Etiqueta	**U**	**S**	**T**	**%**
Pregunta *(Pergunta)*	5.474	864	6.338	27
Respuesta *(Resposta)*	1.839	2.446	4.285	18
Confirmación *(Confirmação)*	0	3.629	3.629	15
Nueva Consulta *(Nova Consulta)*	0	2.474	2.474	11
Espera *(Espera)*	0	1.948	1.948	8
Cierre *(Encerramento)*	927	900	1.827	8
Afirmación *(Aceitação)*	990	0	990	4
Apertura *(Abertura)*	0	900	900	4
No Entendido *(Não Percebido)*	4	653	657	3
Negación *(Rejeição)*	340	0	340	1
Indefinida *(Indefinida)*	141	18	159	1

Tabela 2: Distribuição das etiquetas de Nível 1 no corpus. A tradução em português de cada etiqueta está entre parêntesis. As colunas identificadas como U, S e T referem-se ao número de segmentos de utilizador, sistema e total anotados com a etiqueta.

Embora partilhem a maioria das etiquetas, os dois níveis inferiores da hierarquia focam-se em diferentes tipos de informação específica da tarefa. Enquanto o segundo nível está relacionado com o tipo de informação que é implicitamente focado pelo segmento, o terceiro nível está relacionado com o tipo de informação que é explicitamente referido no segmento. A título ilustrativo, vamos olhar para o segmento *"Estou à procura de comboios de Bilbau para a Corunha na segunda-feira, 16 de Fevereiro de 2004."*, extraído do diálogo da Figura 1. Uma vez que o segmento revela a intenção de encontrar um horário de comboio, este tem *Hora de Partida* como etiqueta de Nível 2. No entanto, como esse horário de partida não é explicitamente referido no segmento, essa etiqueta não faz parte das suas etiquetas de Nível 3. Por outro lado, o segmento refere explicitamente um local de partida, um destino e uma data. Logo, tem as etiquetas de Nível 3 correspondentes — *Origem*, *Destino* e *Dia*.

A distribuição das etiquetas de ambos os níveis é mostrada na Tabela 3. Podemos ver que existem 10 etiquetas comuns e três adicionais no Nível 3 — *Número de Ordem*, *Número de Comboios* e *Tipo de Viagem*. Para além disso, ambos os níveis têm a etiqueta *Nula*, que representa a ausência de etiqueta nesse nível. Neste sentido, podemos ver que apenas 63% dos segmentos têm etiquetas de Nível 2, e que a percentagem é ainda menor, 52%, quando se consideram etiquetas de Nível 3. Isto deve-se principalmente ao facto de que segmentos etiquetados como *Abertura*, *Encerramento*, *Indefinida*, *Não Entendido*, *Espera*, e *Nova Consulta* no primeiro nível não podem ter etiquetas nos restantes níveis. Por fim, é importante referir que cada segmento tem uma e apenas uma etiqueta de Nível 1, mas pode ter várias etiquetas de Nível 2 e Nível 3.

Como observação final, é importante referir que algumas etiquetas de Nível 2 — *Duração*, *Classe* e *Serviço* — e de Nível 3 — *Serviço* e *Duração* — ocorrem apenas em 0,1% dos segmentos ou menos. Por isso, essas etiquetas são especialmente difíceis de prever usando métodos de aprendizagem automática que se focam em maximizar a taxa de acerto no corpus como um todo.

3.2 Arquitectura da Rede

Uma vez que queremos avaliar o desempenho de diferentes abordagens baseadas em DNNs no reconhecimento de actos de diálogo no corpus DIHANA, precisamos de definir uma base comum para comparação. Para tal, usamos uma arquitectura de rede genérica, mostrada na Figura 2, baseada naquelas das abordagens com melhor desempenho referidas na Secção 2.2. Usando esta arquitectura, a abordagem para identificar os actos de diálogo de um segmento é a seguinte: Primeiro, o segmento é dividido em *tokens*, que são passados por uma camada de *embedding* para gerar as suas representações nessa forma. Em seguida, a sequência de *embeddings* é passada para a abordagem de representação do segmento. A representação obtida pode então ser complementada com informação de contexto, antes de ser passada por uma camada de redução de dimensionalidade. Por fim, a representação reduzida é passada para a camada de saída, que gera a classificação de actos de diálogo do segmento. A motivação para cada um destes passos e a forma como as suas características variam de acordo com o nível da hierarquia em foco são descritas

Nível 2					Nível 3				
Etiqueta	**U**	**S**	**T**	**%**	**Etiqueta**	**U**	**S**	**T**	**%**
Nulo *(Nula)*	1.923	6.893	8.816	37	Nulo *(Nula)*	2.954	8.317	11.271	48
Hora Salida *(Hora de Partida)*	3.309	3.523	7.432	32	Destino *(Destino)*	1.631	2.079	3.710	16
Precio *(Preço)*	2.071	1.267	3.338	14	Día *(Dia)*	1.881	1.778	3.659	16
Día *(Dia)*	1.026	923	1.949	8	Origen *(Origem)*	896	2.085	2.981	13
Origen *(Origem)*	477	480	957	4	Hora Salida *(Hora de Partida)*	692	1.633	2.325	10
Destino *(Destino)*	452	400	852	4	Número Trenes *(Número de Comboios)*	0	1.863	1.863	8
Tipo Tren *(Tipo de Comboio)*	317	226	543	2	Tipo Tren *(Tipo de Comboio)*	544	1.253	1.797	8
Hora Llegada *(Hora de Chegada)*	90	88	178	1	Número Orden *(Número de Ordem)*	84	950	1.034	4
Tiempo Recorrido *(Duração)*	14	15	29	0,1	Clase Billete *(Classe)*	129	766	895	4
Clase Billete *(Classe)*	15	12	27	0,1	Precio *(Preço)*	47	731	778	3
Servicio *(Serviço)*	3	5	8	0	Hora Llegada *(Hora de Chegada)*	199	490	689	3
					Tipo Viaje *(Tipo de Viagem)*	643	0	643	3
					Servicio *(Serviço)*	15	4	19	0,1
					Tiempo Recorrido *(Duração)*	0	14	14	0,1

Tabela 3: Distribuição das etiquetas de Nível 2 e Nível 3 no corpus. A tradução em português de cada etiqueta está entre parêntesis. As colunas identificadas como U, S e T referem-se ao número de segmentos de utilizador, sistema e total anotados com a etiqueta.

abaixo.

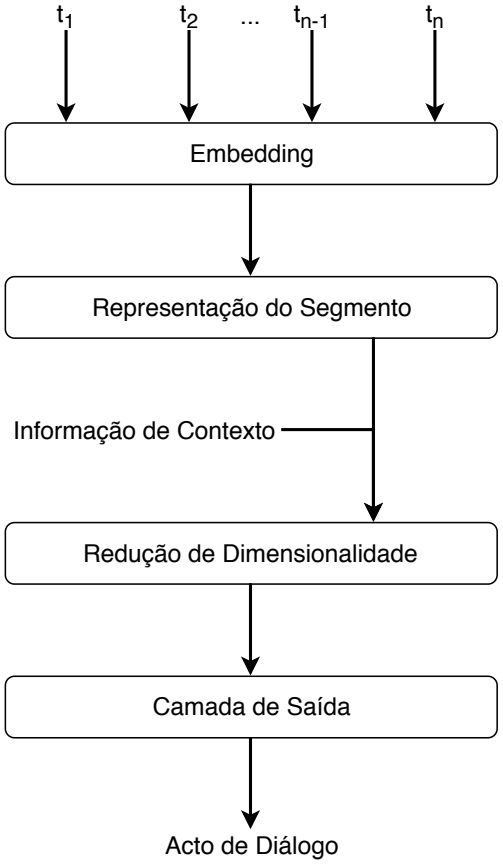

Figura 2: A arquitectura genérica das redes usadas nas nossas experiências. t_i corresponde ao i-ésimo *token*.

3.2.1 Embedding

A entrada da nossa rede é a sequência de *tokens* do segmento. De forma semelhante à maioria dos estudos anteriores sobre o reconhecimento de actos de diálogo, nós utilizamos tokenização no nível da palavra. Como mostrado no nosso estudo anterior (Ribeiro et al., 2018), o nível do carácter também é capaz de fornecer informação importante. No entanto, como forma de simplificação, não o incluímos neste estudo. Para além disso, ignoramos a pontuação, pois esta pode não estar disponível para um sistema de diálogo. Os tokens são então passados para a camada de *embedding* para serem transformados numa representação vectorial correspondente à sua posição no espaço de *embeddings*. Nas nossas experiências usamos *embeddings* pré-treinados usando o método Word2Vec (Mikolov et al., 2013) no Spanish Billion Words Corpus (Cardellino, 2016). Embora tenhamos explorado espaços de *embeddings* com diferentes dimensionalidades, apenas reportamos os resultados obtidos utilizando a dimensionalidade 200, tal como no estudo de Liu et al. (2017), uma vez que esta levou consistentemente a melhores resultados do que as dimensionalidades exploradas por Khanpour et al. (2016).

3.2.2 Representação do Segmento

Este passo gera uma representação vectorial do segmento através da combinação das representações dos seus *tokens*. Tal como referido na Secção 2.2, as abordagens com melhor desempenho no reconhecimento de actos de diálogo em dados em inglês diferem neste passo. Enquanto a abordagem de Khanpour et al. (2016) é baseada em RNNs, a de Liu et al. (2017) é baseada em CNNs. Ambas têm as suas vantagens, uma vez que enquanto a primeira se foca em capturar informação de sequências de *tokens* relevan-

tes, a segunda foca-se no contexto que circunda cada *token* e, por isso, captura padrões relevantes. Uma vez que os diferentes níveis da hierarquia de anotação de actos de diálogo do corpus DIHANA têm diferentes características, nós usamos ambas as abordagens nas nossas experiências para avaliar se existe uma com melhor desempenho em qualquer situação ou se existe uma dependência do nível em foco.

Tal como descrito na Secção 2.2, a abordagem baseada em RNNs de Khanpour et al. (2016) usa uma pilha de 10 unidades LSTM. A representação do segmento é dada pela concatenação das saídas das 10 unidades após processarem todos os *tokens* do segmento. Usar as saídas após o processamento de todos os *tokens* faz sentido, uma vez que estes são processados sequencialmente pelas unidades recorrentes e, por isso, essas saídas contêm informação de todo o segmento. Os resultados reportados neste artigo foram obtidos usando uma pilha de cinco Unidades Recorrentes *Gated* (GRUs) em vez de 10 LSTMs, uma vez que, nas nossas experiências preliminares, o desempenho foi semelhante, mas com um consumo de recursos significativamente menor. A Figura 3 mostra uma representação gráfica desta abordagem.

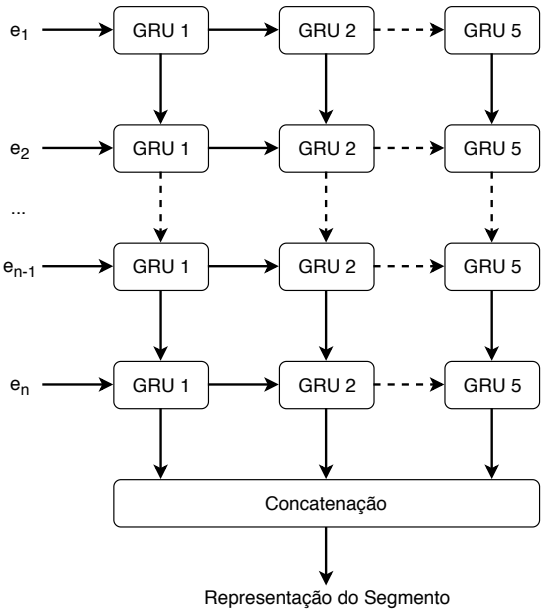

Figura 3: A abordagem de representação do segmento baseada em RNNs. e_i corresponde à representação do i-ésimo *token* na forma de *embedding*.

Também como descrito na Secção 2.2, a abordagem baseada em CNNs de Liu et al. (2017) usa três CNNs temporais paralelas com janelas de contexto com tamanho entre um e três, inclusive. Isto significa que a abordagem se foca em conjuntos de no máximo três palavras consecutivas. Um estudo anterior (Kim, 2014) usou janelas de contexto com tamanho entre três e cinco, de forma a capturar relações entre palavras mais distantes que eram relevantes para as tarefas exploradas. Tendo em conta a tarefa que estamos a explorar, as janelas de contexto mais relevantes dependem do nível em foco, uma vez que os actos de diálogo específicos da tarefa estão tipicamente relacionados com a presença de palavras específicas, enquanto os actos de diálogo genéricos estão mais relacionados com a estrutura do segmento e, consequentemente, com janelas mais largas. Para confirmar isto, usamos os dois conjuntos de janelas de contexto nas nossas experiências. As saídas das CNNs são filtradas usando uma operação de *max pooling* e são em seguida concatenadas para gerar a representação do segmento. A Figura 4 mostra uma representação gráfica desta abordagem.

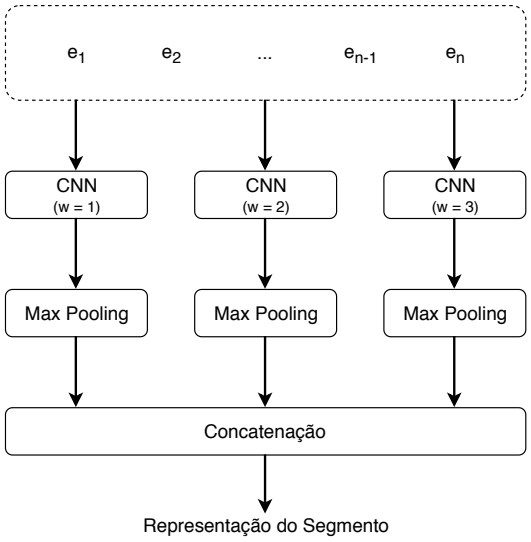

Figura 4: A abordagem de representação do segmento baseada em RNNs. e_i corresponde à representação do i-ésimo *token* na forma de *embedding*. O parâmetro w refere-se ao tamanho da janela de contexto.

3.2.3 Informação de Contexto

Vários estudos anteriores confirmaram a importância de informação de contexto extraída dos segmentos anteriores para a tarefa de reconhecimento de actos de diálogo (Ribeiro et al., 2015; Lee & Dernoncourt, 2016; Liu et al., 2017). Adicionalmente, esses estudos mostraram que a influência dos segmentos anteriores decresce com a distância e que as classificações de actos de

diálogo desses segmentos são mais informativas que as suas palavras. Por isso, nas nossas experiências fornecemos informação de contexto à rede usando a mesma abordagem baseada nas etiquetas dos segmentos anteriores usada no nosso estudo anterior no corpus SwDA (Ribeiro et al., 2015) e por Liu et al. (2017). Isto é, as etiquetas dos segmentos anteriores são transformadas numa representação vectorial única e concatenadas à representação do segmento. Tal como Liu et al. (2017), exploramos o uso de informação de contexto extraída de até um máximo de três segmentos anteriores, uma vez que o nosso estudo anterior mostrou que não existem melhorias significativas ao usar segmentos adicionais. No contexto de um sistema de diálogo que tenta identificar a intenção do seu interlocutor, este só tem acesso aos segmentos anteriores. Como tal, não usamos informação extraída de segmentos futuros nas nossas experiências. É importante referir que usamos as anotações manuais dos segmentos para fornecer a informação de contexto. Portanto, os resultados obtidos representam um tecto para o desempenho da abordagem. Optámos por não usar etiquetas obtidas automaticamente, uma vez que tanto o nosso estudo e o de Liu et al. (2017) mostraram que esta abordagem tem melhor desempenho que aquelas que usam as palavras de segmentos anteriores, mesmo quando as etiquetas são obtidas automaticamente. De acordo com esses estudos, é esperado que a taxa de acerto decresça cerca de dois pontos percentuais ao usar etiquetas obtidas automaticamente. No entanto, um sistema de diálogo está ciente dos actos de diálogo dos seus próprios segmentos. Como tal, só a classificação dos segmentos do utilizador está sujeita a erro, o que se espera que reduza o decréscimo da taxa de acerto. Ainda assim, como trabalho futuro, é importante avaliar qual o valor real desse decréscimo neste cenário.

Adicionalmente, uma vez que o corpus DIHANA está anotado com etiquetas de actos de diálogo hierárquicas, quando nos focamos num dado nível, exploramos também o uso de informação de contexto extraída dos níveis superiores, tanto relativamente ao segmento actual como aos anteriores. Para fornecer essa informação, usamos a mesma abordagem baseada em etiquetas usada para fornecer informação de contexto dos segmentos anteriores.

3.2.4 Redução de Dimensionalidade

Para evitar possíveis diferenças de resultados causadas pelo uso de representações de segmentos com diferente dimensionalidade, a nossa arquitectura inclui uma camada de redução de dimensionalidade que mapeia a representação do segmento, incluindo informação de contexto, num espaço com 100 dimensões. Deste modo, as diferenças de desempenho que possam ser observadas devem-se à natureza da abordagem de representação do segmento e à informação que esta é capaz de capturar e não a factores relacionados com a dimensionalidade. Para além disso, para reduzir a probabilidade de haver um sobreajustamento aos dados de treino, esta camada aplica também uma técnica de *dropout*, desactivando 50% dos neurónios durante a fase de treino.

3.2.5 Camada de Saída

A camada de saída mapeia a representação reduzida do segmento nas etiquetas de actos de diálogo correspondentes. Este processo é feito usando uma camada totalmente ligada com um número de neurónios igual ao número de etiquetas. Como cada segmento tem apenas uma etiqueta de Nível 1, usamos *softmax* como função de activação e a entropia cruzada categórica como função de custo. No entanto, essa abordagem não é válida para os restantes níveis, uma vez que estes permitem que um segmento tenha múltiplas etiquetas. Por isso, nesses casos, usamos a função de activação sigmoide e a entropia cruzada binária como função de custo que, tendo em conta a possibilidade de múltiplas etiquetas, é na verdade a função de custo de Hamming, apropriada para este tipo de problema (Díez et al., 2015). Em ambos os casos, por questões de desempenho, usamos o optimizador Adam (Kingma & Ba, 2015).

3.3 Treino e Avaliação

Para implementar as nossas redes usámos a API de alto nível Keras (Chollet et al., 2015) fornecida com a biblioteca TensorFlow (Abadi et al., 2015). Usámos uma metodologia de treino em lotes de tamanho 512 e interrompemos o treino após 10 épocas sem melhorias no conjunto de validação. Uma vez que existe algum não-determinismo envolvido, especialmente devido à execução em Unidade de Processamento Gráfico (GPU), os resultados apresentados na próxima secção referem-se à média (m) e ao desvio padrão (s) dos resultados obtidos em 10 execuções.

Para avaliar as nossas abordagens, fizemos uma validação cruzada com cinco partições, usando as partições definidas nos primeiros estudos sobre o corpus DIHANA (Tamarit & Martínez-Hinarejos, 2008; Martínez-Hinarejos et al., 2008). As métricas de avaliação utilizadas

variam de acordo com o nível da hierarquia em foco. Uma vez que cada segmento tem apenas uma etiqueta de Nível 1, neste caso lidamos com um problema de classificação de etiqueta única. Como tal, de forma semelhante a estudos anteriores sobre reconhecimento automático de actos de diálogo, avaliamos o desempenho usando a taxa de acerto (Acc). No entanto, essa não é a métrica mais adequada para os Níveis 2 e 3 da hierarquia, uma vez que estes colocam problemas de classificação multi-etiqueta. Por isso, avaliamos o desempenho nesses níveis usando as métricas adaptadas a cenários multi-etiqueta descritas por Sorower (2010). A métrica equivalente à taxa de acerto em cenários multi-etiqueta é a taxa de correspondência exacta (MR), definida como

$$\mathrm{MR} = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = Z_i), \tag{1}$$

onde Y_i é o conjunto de etiquetas de referência do exemplo i, Z_i é o conjunto de etiquetas previstas pelo classificador para o mesmo exemplo e I é a função indicadora. O problema desta métrica é que não considera acertos parciais, que são comuns em problemas de classificação multi-etiqueta. De forma a considerar esses casos, as métricas tradicionais para problemas de etiqueta única — taxa de acerto (Acc), precisão (P), sensibilidade (R) e medida-F (F_1) — são adaptadas da seguinte forma:

$$\mathrm{Acc} = \frac{1}{n}\sum_{i=1}^{n} \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|}, \tag{2}$$

$$P = \frac{1}{n}\sum_{i=1}^{n} \frac{|Y_i \cap Z_i|}{|Z_i|}, \tag{3}$$

$$R = \frac{1}{n}\sum_{i=1}^{n} \frac{|Y_i \cap Z_i|}{|Y_i|}, \tag{4}$$

$$F_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{2|Y_i \cap Z_i|}{|Y_i| + |Z_i|}. \tag{5}$$

onde o operador $|X|$ é usado para obter a cardinalidade do conjunto X. Para além disso, tal como referido na Secção 3.2.5, a função de custo de Hamming (HL), que diz quantas vezes, em média, a relevância de um exemplo para uma etiqueta é incorrectamente prevista e é definida como

$$\mathrm{HL} = \frac{1}{n|L|}\sum_{i=1}^{n}\sum_{l\in L}[I(l \in Z_i \wedge l \notin Y_i) + I(l \notin Z_i \wedge l \in Y_i)], \tag{6}$$

onde L é o conjunto de todas as etiquetas, é também uma métrica apropriada para avaliar problemas de classificação multi-etiqueta. Na próxima secção, os resultados de todas as métricas excepto a função de custo de Hamming serão apresentados na forma de percentagens.

Para verificar se as diferenças entre os resultados de duas abordagens são estatisticamente significativas, escolhemos aleatoriamente uma das execuções de cada uma das abordagens e aplicámos um teste binomial sobre a sua taxa de acerto, no caso das experiências sobre o Nível 1, e sobre a sua taxa de correspondência exacta, no caso das experiências sobre os Níveis 2 e 3. Ao longo da discussão apresentada na próxima secção, consideramos um nível de confiança de 95%, isto é, consideramos que existe uma diferença estatisticamente significativa entre duas abordagens se o valor-p do teste binomial for inferior a 0,05.

4 Resultados

Uma vez que cada nível da anotação hierárquica de actos de diálogo do corpus DIHANA tem características diferentes e coloca problemas de diferentes tipos, começamos por apresentar os resultados alcançados em cada um dos níveis de forma independente. Para além disso, como queremos avaliar a importância da informação de contexto dos níveis superiores, começamos no nível superior e descemos na hierarquia. Por fim, apresentamos os resultados obtidos na combinação hierárquica dos diferentes níveis.

4.1 Nível 1

Os resultados obtidos ao usar as duas abordagens de representação de segmento para prever as etiquetas de Nível 1 são mostrados na Tabela 4. Podemos ver que a abordagem baseada em CNNs tem melhor desempenho que a baseada em RNNs ($p \approx 0{,}04$). No entanto, ambas levam a uma taxa de acerto superior a 90% e a diferença entre elas é de apenas 0,5 pontos percentuais, o que sugere que a informação sobre intenção que são capazes de capturar é semelhante. No entanto, enquanto o treino da rede da abordagem baseada em CNNs demora em média 0,61 segundos por época e necessita de cerca de 27 épocas para convergir, o treino da rede da abordagem baseada em RNNs demora muito mais tempo, com uma média 17,63 segundos por época e 46 épocas para convergir.

Adicionalmente, tal como esperado, usar CNNs com janelas de contexto mais largas leva a melhores resultados ($p \approx 0{,}03$), o que confirma

Abordagem	Acc m	Acc s
Recorrente (RNN)	91,20	0,06
Convolucional (CNN) w = [1,3]	91,46	0,12
Convolucional (CNN) w = [3,5]	91,70	0,13

Tabela 4: Taxa de acerto nas etiquetas de Nível 1 usando as duas abordagens de representação de segmento.

que as etiquetas genéricas do Nível 1 estão mais relacionadas com a estrutura do diálogo do que com palavras específicas. Ainda assim, uma vez que usamos três CNNs paralelas e existe sobreposição entre os dois conjuntos de janelas usados nas nossas experiências, a diferença em termos de taxa de acerto entre usar as janelas mais estreitas usadas por Liu et al. (2017) e as mais largas usadas por Kim (2014) é de apenas 0,24 pontos percentuais.

Relativamente à informação de contexto fornecida pelos segmentos anteriores, os resultados na Tabela 5 mostram que o primeiro segmento anterior é o mais importante, levando a uma melhoria da taxa de acerto na ordem dos 4,45 pontos percentuais ($p \approx 6{,}7e^{-167}$). Uma melhoria adicional de 1,77 pontos percentuais é alcançada fornecendo informação de dois segmentos adicionais ($p \approx 4{,}6e^{-58}$). Este padrão era esperado, uma vez que já tinha sido observado tanto no nosso estudo (Ribeiro et al., 2015) como no de Liu et al. (2017) no corpus SwDA, que também está anotado com etiquetas de actos de diálogo genéricas e independentes do domínio e da tarefa.

n	Acc m	Acc s
0	91,70	0,13
1	96,15	0,08
2	97,47	0,06
3	97,92	0,04

Tabela 5: Taxa de acerto nas etiquetas de Nível 1 usando informação de contexto de n segmentos anteriores.

Quando é utilizada informação de contexto de três segmentos anteriores, o classificador só falha a previsão de dois por cento dos segmentos. Este resultado tem em conta todos os segmentos do corpus DIHANA. No entanto, os segmentos do sistema são estruturados à priori e, portanto, são mais fáceis de prever do que os segmentos do utilizador. De facto, se considerarmos um cenário em que um sistema de diálogo tenta prever actos de diálogo, ele está ciente dos seus próprios actos e tem apenas de prever os dos seus interlocutores. Neste sentido, na Tabela 6 mostramos os resultados obtidos ao considerar os segmentos do utlizador e do sistema independentemente. Como esperado, a taxa de acerto média nos segmentos do sistema é de 99,91%. Nos segmentos do utilizador esse valor diminui para 95,17%, o que ainda assim revela um elevado desempenho.

Orador	Acc m	Acc s
Utilizador	95.17	0.12
Sistema	99.91	0.00

Tabela 6: Taxa de acerto nas etiquetas de Nível 1 em segmentos do utilizador e do sistema.

Olhando para cada etiqueta individualmente, a mais difícil de identificar é a *Indefinida*, com uma sensibilidade de cerca de 57%. Isto era esperado, uma vez que essa etiqueta cobre todos os casos que não podem ser etiquetados com nenhuma das outras etiquetas, incluindo problemas no diálogo. Para todas as etiquetas restantes, o valor da sensibilidade é acima de 95%, sendo o mais baixo o da etiqueta *Resposta*, que é também aquela com valor mais baixo em termos de precisão (96%). Em ambos os casos, a confusão é tipicamente com a etiqueta *Pergunta*, o que faz sentido, uma vez que perguntas e respostas podem ter as mesmas palavras e diferir apenas em termos da sua ordem. De facto, se considerarmos questões em forma declarativa, pode não haver qualquer tipo de diferença.

Considerando estudos anteriores sobre reconhecimento de actos de diálogo no corpus DIHANA, só Tamarit & Martínez-Hinarejos (2008) é que avaliaram o desempenho no Nível 1 individualmente, atingindo uma taxa de acerto de 60,70%. No entanto, o estudo focava-se no uso de informação prosódica e, como tal, não é justo comparar os resultados com os nossos, pois a nossa abordagem tira partido das transcrições.

4.2 Level 2

Tal como referido na Secção 3.1, algumas etiquetas de Nível 1 só podem ser emparelhadas com a etiqueta *Nula* nos restantes níveis. Logo, os segmentos anotados com uma dessas etiquetas no Nível 1, ficam com as etiquetas dos restantes níveis definidas automaticamente, independentemente do seu conteúdo. Por isso, esses segmentos

Abordagem	MR		Acc		P		R		F_1		HL	
	m	s	m	s	m	s	m	s	m	s	m	s
RNN	69,65	0,50	70,42	0,48	71,10	0,46	70,51	0,47	70,68	0,47	0,0381	0,0004
CNN w = [1,3]	70,71	0,33	71,58	0,33	72,30	0,33	71,74	0,34	71,87	0,33	0,0381	0,0002
CNN w = [3,5]	70,24	0,27	71,17	0,26	71,93	0,28	71,33	0,26	71,48	0,26	0,0383	0,0000

Tabela 7: Resultados obtidos no Nível 2 usando as duas abordagens de representação de segmento.

	MR		Acc		P		R		F_1		HL	
n	m	s	m	s	m	s	m	s	m	s	m	s
0	70,71	0,33	71,58	0,33	72,30	0,33	71,74	0,34	71,87	0,33	0,0381	0,0002
1	91,07	0,14	91,52	0,13	91,84	0,13	91,67	0,13	91,68	0,13	0,0121	0,0002
2	92,52	0,09	92,99	0,08	93,30	0,08	93,12	0,09	93,14	0,08	0,0101	0,0001
3	92,97	0,12	93,45	0,11	93,75	0,09	93,61	0,11	93,60	0,10	0,0094	0,0001

Tabela 8: Resultados obtidos no Nível 2 usando informação de contexto de n segmentos anteriores.

	MR		Acc		P		R		F_1		HL	
n	m	s	m	s	m	s	m	s	m	s	m	s
0	93,18	0,18	93,68	0,16	93,99	0,15	93,87	0,15	93,63	0,15	0,0092	0,0002
1	94,28	0,15	94,75	0,14	95,06	0,13	94,91	0,13	94,91	0,13	0,0077	0,0002
2	94,29	0,05	94,76	0,05	95,06	0,05	94,91	0,06	94,91	0,06	0,0077	0,0001
3	94,38	0,11	94,84	0,11	95,15	0,12	94,97	0,12	94,99	0,12	0,0075	0,0001

Tabela 9: Resultados obtidos no Nível 2 usando informação de Nível 1 de n segmentos anteriores.

Orador	MR		Acc		P		R		F_1		HL	
	m	s	m	s	m	s	m	s	m	s	m	s
Utilizador	91,28	0,24	92,08	0,21	92,62	0,19	92,32	0,22	92,34	0,21	0,0115	0,0003
Sistema	98,43	0,09	98,44	0,08	98,44	0,08	98,45	0,07	98,44	0,08	0,0024	0,0001

Tabela 10: Resultados obtidos no Nível 2 em segmentos do utilizador e do sistema.

não são considerados nas nossas experiências sobre os Níveis 2 e 3.

De forma semelhante ao que observámos para o Nível 1, na Tabela 7 podemos ver que usar a abordagem de representação de segmento baseada em CNNs leva a melhores resultados do que a baseada em RNNs. A única excepção é o resultado ao nível da função de custo de Hamming que, em média, é igual para ambas as abordagens. Em todas as outras métricas, a abordagem baseada em CNNs supera a baseada em RNNs por mais de um ponto percentual ($p \approx 1{,}1e^{-14}$). No entanto, neste caso, a discrepância no número de épocas de treino necessárias para haver convergência é menor, com uma média de 46 para a abordagem baseada em CNNs e 56 para a baseada em RNNs. Para além disso, uma vez que são considerados menos segmentos, os tempos de treino por época são reduzidos para 0,40 e 11,67 segundos, respectivamente.

Por outro lado, contrariamente ao observado para o Nível 1, usar janelas de contexto mais estreitas parece levar a melhores resultados. No entanto, a diferença não é estatisticamente significativa ($p \approx 0{,}12$). Ainda assim, isto mostra que as etiquetas dependentes do domínio estão mais relacionadas com palavras específicas do que as etiquetas genéricas do Nível 1. Para além disso, uma vez que o número de etiquetas por segmento é tipicamente baixo, os classificadores tendem a evitar escolher etiquetas incorrectas, o que se reflecte numa precisão mais alta do que a sensibilidade em todas as abordagens.

Os resultados mostrados na Tabela 8 mostram que, de forma semelhante ao que acontece no Nível 1, os segmentos anteriores fornecem informação de contexto relevante para a tarefa. No entanto, neste caso, a importância do primeiro segmento anterior é mais pronunciada, levando a uma redução do custo para menos de um terço e

melhorando as restantes métricas em cerca de 20 pontos percentuais ($p \approx 5{,}0e^{-324}$). Isto faz sentido, considerando que os diálogos incluem uma grande quantidade de pares pergunta-resposta focados no mesmo tipo de informação, que é o foco das etiquetas de Nível 2. Logo, nesses casos, as etiquetas de Nível 2 dos dois segmentos são as mesmas e, por isso, as etiquetas do primeiro segmento anterior fornecem uma pista importante para a identificação das etiquetas do segmento actual.

Na Tabela 9, podemos ver que informação extraída do Nível 1 também é importante. Usar informação do segmento actual leva a uma melhoria que, embora significativa ($p \approx 0{,}01$), é apenas na ordem dos 0,2 pontos percentuais. No entanto, considerar também a etiqueta de Nível 1 do segmento anterior leva a uma melhoria na ordem dos 1,5 pontos percentuais ($p \approx 8{,}7e^{-6}$). Isto continua a ser explicado pela presença de um grande número de pares pergunta-resposta nos diálogos, uma vez que se o segmento anterior estiver etiquetado como *Pergunta* no Nível 1, então é provável que o segmento actual tenha as mesmas etiquetas de Nível 2 que esse segmento. Utilizar informação extraída de segmentos adicionais não leva a melhorias significativas ($p \approx 0{,}74$).

De forma semelhante ao que observámos para o Nível 1, o desempenho nos segmentos do sistema é diferente do nos segmentos do utilizador. Na Tabela 10 podemos ver que nos segmentos do sistema, os resultados ao nível de todas as métricas percentuais rondam os 98,4%, enquanto nos segmentos do utilizador a taxa de correspondência exacta é de 91,28% e as restantes métricas percentuais rondam os 92%.

Considerando as etiquetas individualmente, a melhor abordagem não é capaz de identificar nenhuma das três etiquetas menos predominantes no corpus. No entanto, isto era esperado, uma vez que nenhuma delas ocorre em mais de 29 segmentos. Como tal, elas são irrelevantes do ponto de vista de uma abordagem focada em reduzir o erro no corpus como um todo e necessitam de abordagens especializadas ou de mais dados para serem identificadas. O valor da medida-F para a etiqueta *Hora de Chegada* é de cerca de 75%, uma vez que esta é facilmente confundível com a etiqueta *Hora de Partida* e é a menos predominante das duas. Embora a precisão da etiqueta *Tipo de Comboio* seja acima de 95%, o valor da sensibilidade para a mesma etiqueta é de apenas 87%. Isto acontece porque a etiqueta aparece em apenas 2% dos segmentos. Como tal, em segmentos que se focam em múltiplos aspectos, a informação das palavras relacionadas com o tipo de comboio é descartada em favor de informação que permita identificar as etiquetas mais predominantes. Todas as etiquetas restantes têm um valor de medida-F acima dos 95% com balanço entre a precisão e a sensibilidade.

Os estudos anteriores sobre o reconhecimento de actos de diálogo no corpus DIHANA não exploraram o Nível 2 individualmente, mas sim em combinação com o Nível 1, usando a combinação das etiquetas dos dois níveis como conjunto de etiquetas e abordando o problema como um problema de classificação de etiqueta única semelhante ao colocado pelo Nível 1. Logo, os nossos resultados no Nível 2 não podem ser comparados directamente com os desses estudos. Os resultados obtidos na combinação dos dois níveis são discutidos na Secção 4.4.

4.3 Nível 3

A Tabela 11 mostra que, de forma semelhante ao que observámos nos restantes níveis, usar a abordagem de representação de segmento baseada em CNNs leva a melhores resultados do que a baseada em RNNs ($p \approx 9{,}6e^{-5}$). No entanto, neste caso a diferença é menos pronunciada. De facto, ao usar o conjunto de janelas de contexto mais largas, a abordagem baseada em CNNs tem pior desempenho que a baseada em RNNs ($p \approx 1{,}2e^{-4}$). Isto deve-se ao facto de o Nível 3 se focar na informação que é referida explicitamente nos segmentos e, como tal, ser ainda mais orientado a palavras específicas do que o Nível 2. Este facto também explica que, em média, os resultados de todas as métricas percentuais sejam superiores a 96%. Os tempos médios por época são iguais aos registados para o Nível 2. No entanto, são necessárias mais épocas para atingir convergência — 86 para a abordagem baseada em RNNs e 80 para a baseada em CNNs.

Os resultados da Tabela 12 mostram que, neste caso, a melhoria obtida ao usar informação dos segmentos anteriores ao mesmo nível é desprezável e não é estatisticamente significativa ($p \approx 0{,}48$). Uma vez mais, isto pode ser explicado pela natureza do Nível 3 e o seu foco no que é referido explicitamente no segmento actual. Logo, informação relativa ao que é referido explicitamente nos segmentos anteriores não é relevante.

Na Tabela 13 podemos ver que a informação fornecida pelo Nível 2 é ligeiramente superior à fornecida pelas etiquetas de Nível 3 dos segmentos anteriores. Neste caso, considerar a etiqueta de Nível 2 do mesmo segmento leva uma melhoria estatisticamente significativa ($p \approx 0{,}03$). Esta melhoria pode ser explicada pelo facto de que

Abordagem	MR		Acc		P		R		F_1		HL	
	m	s	m	s	m	s	m	s	m	s	m	s
RNN	95,79	0,24	96,61	0,29	96,84	0,29	96,81	0,30	96,78	0,30	0,0043	0,0004
CNN w = [1,3]	96,01	0,08	96,88	0,10	97,11	0,10	97,08	0,12	97,05	0,11	0,0040	0,0000
CNN w = [3,5]	95,35	0,23	96,26	0,18	96,51	0,17	96,45	0,15	96,44	0,16	0,0046	0,0002

Tabela 11: Resultados obtidos no Nível 3 usando as duas abordagens de representação de segmento.

n	MR		Acc		P		R		F_1		HL	
	m	s	m	s	m	s	m	s	m	s	m	s
0	96,01	0,08	96,88	0,10	97,11	0,10	97,08	0,12	97,05	0,11	0,0040	0,0000
1	96,05	0,13	96,91	0,10	97,14	0,09	97,10	0,09	97,08	0,10	0,0039	0,0001
2	96,10	0,16	96,95	0,11	97,17	0,11	97,14	0,10	97,12	0,10	0,0039	0,0002
3	96,10	0,16	96,96	0,13	97,19	0,13	97,14	0,11	97,13	0,12	0,0039	0,0001

Tabela 12: Resultados obtidos no Nível 3 usando informação de contexto de n segmentos anteriores.

n	MR		Acc		P		R		F_1		HL	
	m	s	m	s	m	s	m	s	m	s	m	s
0	96,20	0,15	97,03	0,11	97,25	0,11	97,20	0,10	97,19	0,11	0,0037	0,0002
1	96,24	0,09	97,05	0,08	97,28	0,08	97,22	0,07	97,21	0,08	0,0037	0,0000
2	96,29	0,08	97,11	0,08	97,34	0,09	97,27	0,09	97,26	0,09	0,0036	0,0000
3	96,17	0,06	97,00	0,06	97,23	0,06	97,18	0,06	97,17	0,06	0,0038	0,0000

Tabela 13: Resultados obtidos no Nível 3 usando informação de Nível 2 de n segmentos anteriores.

n	MR		Acc		P		R		F_1		HL	
	m	s	m	s	m	s	m	s	m	s	m	s
0	96,32	0,12	97,13	0,10	97,36	0,09	97,29	0,10	97,28	0,09	0,0036	0,0001
1	96,29	0,14	97,10	0,12	97,33	0,12	97,26	0,11	97,26	0,12	0,0037	0,0001
2	96,34	0,12	97,14	0,11	97,36	0,10	97,30	0,11	97,30	0,11	0,0036	0,0001
3	96,30	0,13	97,13	0,11	97,35	0,10	97,31	0,10	97,29	0,10	0,0037	0,0001

Tabela 14: Resultados obtidos no Nível 3 usando informação de Nível 1 de n segmentos anteriores.

Orador	MR		Acc		P		R		F_1		HL	
	m	s	m	s	m	s	m	s	m	s	m	s
Utilizador	95,58	0,16	95,62	0,17	95,62	0,17	95,65	0,18	95,63	0,17	0,0044	0,0002
Sistema	97,55	0,12	99,06	0,06	99,52	0,06	99,25	0,03	99,33	0,04	0,0024	0,0001

Tabela 15: Resultados obtidos no Nível 3 em segmentos do utilizador e do sistema.

quando um tipo de informação é referido explicitamente num segmento, tipicamente este também é focado pelo mesmo segmento e, por isso, é comum haver sobreposição das etiquetas dos Níveis 2 e 3. Considerar informação de Nível 2 de segmentos adicionais não leva a melhorias estatisticamente significativas ($p \approx 0{,}13$).

Uma vez que as etiquetas de Nível 1 estão relacionadas com a intenção genérica por trás do segmento, elas não têm relação directa com que é explicitamente referido no segmento e, portanto, com as etiquetas de Nível 3. Isto é confirmado pelos resultados da Tabela 14, que mostram que a melhoria obtida ao utilizar informação do Nível 1 é desprezável e não é estatisticamente significativa ($p \approx 0{,}13$).

Na Tabela 15, podemos ver que, neste caso, a diferença de desempenho nos segmentos do utilizador e do sistema não é tão pronunciada. Mais uma vez, isso é explicado pelo facto de que o Nível 3 é altamente focado em palavras específicas e, portanto, o facto de os segmentos do sistema se-

rem estruturados à priori não tem a mesma influência na classificação.

Considerando as etiquetas individualmente, à semelhança do que acontece no Nível 2, a melhor abordagem é incapaz de identificar as etiquetas menos predominantes, *Duração* e *Serviço*, uma vez que nenhuma delas aparece em mais de 19 segmentos. Das restantes, a etiqueta *Hora de Chegada* é aquela com menor valor de sensibilidade, 88%, uma vez que é facilmente confundível com a mais predominante *Hora de Partida*. Todas as etiquetas restantes têm um valor de medida-F acima de 97% com balanço entre precisão e sensibilidade.

Tal como o Nível 2, os estudos anteriores sobre o reconhecimento de actos de diálogo no corpus DIHANA não exploraram o Nível 3 individualmente, mas sim em combinação com os restantes níveis. Consequentemente, também não é possível comparar directamente os nossos resultados no Nível 3 com os desses estudos. A combinação hierárquica dos vários níveis é explorada na próxima secção.

4.4 Classificação Hierárquica

Tal como referido anteriormente, os estudos anteriores sobre reconhecimento de actos de diálogo no corpus DIHANA não exploraram os níveis específicos da tarefa independentemente, mas sim em combinação com os níveis acima. Isto faz sentido dum ponto de vista hierárquico, uma vez que, supostamente, cada nível é dependente dos que estão acima dele. No entanto, tal como discutido na Secção 3.1, uma vez que cada nível se foca num aspecto diferente relativo à intenção do orador, a única restrição imposta pelo esquema de anotação é que segmentos anotados com uma etiqueta de Nivel 1 relacionada com a estrutura do diálogo ou problemas de comunicação não podem ter etiquetas nos restantes níveis. Ainda assim, os resultados reportados nas secções anteriores mostram que a capacidade de prever as etiquetas de um determinado nível aumenta quando é usada informação de contexto extraída do nível directamente acima. Para além disso, para identificar correctamente a intenção do seu interlocutor, um sistema de diálogo tem de ser capaz de prever correctamente as etiquetas dos três níveis em conjunto. Por isso, também avaliamos o desempenho das nossas abordagens na combinação hierárquica dos vários níveis.

Os estudos anteriores sobre a tarefa abordaram o problema da classificação combinada dos diferentes níveis como um problema de classificação de etiqueta única, em que cada combinação de etiquetas presentes no corpus é considerada uma única etiqueta independente. No entanto, esta abordagem apresenta duas falhas. Por um lado, trata-se de uma simplificação do problema, na medida em que limita as etiquetas possíveis às combinações existentes no corpus. Por outro lado, não tem em conta a natureza multi-etiqueta dos níveis específicos da tarefa.

Contrariamente a esses estudos, nós abordamos o problema hierarquicamente, combinando os melhores classificadores para cada nível. Ou seja, para cada segmento, começamos por prever a sua etiqueta de Nível 1 usando o classificador baseado em CNNs com janelas de contexto mais largas e informação de contexto extraída de três segmentos anteriores. Em seguida, prevemos as suas etiquetas de Nível 2 usando o classificador baseado em CNNs com janelas de contexto mais estreitas, informação de contexto de Nível 2 de três segmentos anteriores e informação de contexto de Nível 1 do segmento actual e do anterior. Por fim, prevemos as suas etiquetas de Nível 3 usando o classificador baseado em CNNs com janelas de contexto mais estreitas e informação de contexto de Nível 2 extraída do segmento actual. De forma a ter em conta o facto de os classificadores dos Níveis 2 e 3 não terem sido treinados nos segmentos com etiquetas de Nível 1 que não permitem etiquetas nos restantes níveis, se o classificador de Nível 1 prevê uma dessas etiquetas para o segmento, os restantes níveis são automaticamente classificados como não tendo etiquetas.

Usando esta abordagem hierárquica, os níveis inferiores ainda são considerados problemas de classificação multi-etiqueta. Logo, todas as combinações de etiquetas são possíveis e não apenas aquelas que aparecem no corpus. Ainda assim, para confirmar que o problema abordado pelos estudos anteriores é realmente mais simples, apresentamos também os resultados alcançados quando a tarefa é abordada como um problema de classificação de etiqueta única. Para obter esses resultados, usámos um classificador com a mesma arquitectura que o melhor classificador de Nível 1, ou seja, um classificador baseado em CNNs com janelas de contexto mais largas e informação de contexto extraída de três segmentos anteriores. No entanto, neste caso, o classificador foi treinado para prever a combinação de todas as etiquetas do segmento de uma só vez e cada uma dessas combinações é vista como uma etiqueta independente.

Para comparação com os resultados obtidos em estudos anteriores, utilizamos a taxa de correspondência exata para avaliar o desempenho quer da abordagem hierárquica, quer da de eti-

queta única. Portanto, se a previsão da etiqueta de Nível 1 for incorrecta ou se houver alguma etiqueta de Nível 2 ou 3 em falta ou adicional, toda a previsão para o segmento é considerada errada.

A Tabela 16 mostra os resultados obtidos na combinação dos Níveis 1 e 2. Usando a abordagem hierárquica, obtivemos, em média, 94,28% de taxa de correspondência exacta, um resultado já acima dos 93,40% reportados por Martínez-Hinarejos et al. (2008) ($p \approx 3{,}0e^{-8}$) e em linha com os 94,08% reportados por Gambäck et al. (2011) ($p \approx 0{,}20$). Ao abordar a tarefa como um problema de classificação de etiqueta única, obtivemos 96,24% de taxa de correspondência exacta, um resultado que é quase dois pontos percentuais acima do resultado obtido usando a abordagem hierárquica ($p \approx 7{,}0e^{-43}$). Isto confirma que a visão do problema como um problema de classificação de etiqueta única é realmente uma simplificação.

Abordagem	**MR** m	s
Hierárquica	94,28	0,03
Etiqueta Única	96,24	0,06
Martínez-Hinarejos et al. (2008)	93,40	
Gambäck et al. (2011)	94,08	

Tabela 16: Resultos obtidos na combinação dos Níveis 1 e 2.

A Tabela 17 mostra os resultados obtidos na combinação dos três níveis. Podemos ver que a maioria das conclusões tiradas para a combinação dos Níveis 1 e 2 também pode ser tirada neste caso. Usando a abordagem hierárquica, obtivemos, em média, uma taxa de correspondência exacta de 92,34 %, que está acima dos 89,70% reportados por Martínez-Hinarejos et al. (2008) e dos 90,97% reportados por Gambäck et al. (2011). No entanto, enquanto na combinação dos dois níveis superiores o resultado da abordagem hierárquica não é estatisticamente diferente do reportado por Gambäck et al. (2011), neste caso há uma melhoria estatisticamente significativa de 1,37 pontos percentuais ($p \approx 6{,}6e^{-14}$). Ao abordar a tarefa como um problema de classificação de etiqueta única, a taxa de correspondência exacta é melhorada para 93,98% ($p \approx 1{,}5e^{-22}$), confirmando uma vez mais que o problema é mais simples.

Abordagem	**MR** m	s
Hierárquica	92,34	0,04
Etiqueta Única	93,98	0,19
Martínez-Hinarejos et al. (2008)	89,70	
Gambäck et al. (2011)	90,97	

Tabela 17: Resultos obtidos na combinação de todos os níveis.

5 Conclusões

Neste artigo explorámos o reconhecimento automático de actos de diálogo no corpus DIHANA. Este corpus e o seu esquema de anotação em três níveis colocam problemas que não têm sido explorados desde que os estudos sobre o reconhecimento de actos de diálogo começaram a focar-se em dados em inglês e, especialmente, no corpus SwDA. O primeiro problema diz respeito à diferença de língua, uma vez que o espanhol tem características diferentes do inglês. Adicionalmente, ao contrário do problema de classificação plana e de etiqueta única colocado pelas anotações SWBD-DAMSL do corpus SwDA, as anotações de actos de diálogo do corpus DIHANA colocam um problema de classificação hierárquica. Para além disso, os dois níveis inferiores dessa hierarquia colocam problemas de classificação multi-etiqueta. Por isso, estudámos como as melhores abordagens para o reconhecimento de actos de diálogo em dados em inglês podem ser aplicadas a estes problemas e quais os aspectos dessas abordagens que são relevantes para a previsão das etiquetas de cada nível, de acordo com suas características.

Uma conclusão comum a todos os níveis é que a abordagem de representação do segmento baseada em CNNs leva a um melhor desempenho do que a abordagem baseada em RNNs. Esta abordagem, aplicada ao reconhecimento de actos de diálogo em dados em inglês por Liu et al. (2017), apresenta três CNNs temporais paralelas com janelas de contexto de diferentes tamanhos. Desta forma, a abordagem de representação do segmento tem em conta conjuntos de palavras de diferentes tamanhos e, dependendo dos tamanhos das janelas, é capaz de capturar informação referente quer a palavras específicas, quer à estrutura do segmento. Nesse sentido, as etiquetas genéricas e independentes da tarefa do Nível 1 estão mais relacionadas com a estrutura do segmento e, por isso, os melhores resultados foram obtidos utilizando um conjunto de janelas de contexto mais largas. Por outro lado, as etiquetas es-

pecíficas da tarefas dos Níveis 2 e 3 estão mais relacionadas com palavras específicas e, por isso, a utilização de um conjunto de janelas mais estreitas levou a um melhor desempenho. Seleccionar um conjunto de janelas adequado é especialmente relevante para prever as etiquetas de Nível 3, uma vez que, ao usar janelas mais largas, a abordagem baseada em CNNs teve um desempenho pior do que a baseada em RNNs. No entanto, isso é explicável pela natureza desse nível, que se foca no tipo de informação que é explicitamente referido nos segmentos e, portanto, a classificação de um segmento é dada pela presença de palavras específicas.

A relação entre as etiquetas de Nível 3 e a presença de palavras específicas no segmento explica também o facto de a informação de contexto extraída dos segmentos anteriores não ser relevante para a previsão dessas etiquetas. Por outro lado, essa informação é relevante para prever as etiquetas dos restantes níveis. No Nível 1, as experiências revelaram um padrão semelhante ao observado tanto no nosso estudo anterior (Ribeiro et al., 2015), como no de Liu et al. (2017) no corpus SwDA, que também está anotado com etiquetas genéricas e independentes da tarefa. No entanto, a importância da informação de contexto dos segmentos anteriores foi especialmente pronunciada nas experiências sobre o Nível 2, reduzindo o valor da função de custo de Hamming para menos de um terço e melhorando as restantes métricas em mais de 20 pontos percentuais. O Nível 2 foca-se no tipo de informação implicitamente focada pelo segmento. Logo, uma vez que os diálogos no corpus DIHANA apresentam múltiplos pares de segmentos focados no mesmo tipo de informação, os segmentos anteriores, especialmente o primeiro, fornecem uma pista importante para a classificação do segmento actual.

Ainda considerando o Nível 2 e as características dos diálogos, a maioria dos pares de segmentos focados no mesmo tipo de informação são pares pergunta-resposta. *Pergunta* e *Resposta* são etiquetas de Nível 1. Por isso, a informação de contexto de Nível 1 extraída quer do segmento actual, quer dos anteriores, também fornece pistas para a previsão de etiquetas de Nível 2. Por outro lado, essa informação é irrelevante para prever etiquetas de Nível 3. No entanto, existe uma relação entre o tipo de informação que é implicitamente focada num segmento e aquela que é explicitamente mencionada nele. Logo, tipicamente, existe sobreposição entre os conjuntos de etiquetas de Nível 2 e 3 de um segmento. Consequentemente, usar informação de contexto extraída do Nível 2 leva a ligeiras melhorias no desempenho ao prever etiquetas de Nível 3.

No corpus DIHANA, os segmentos do sistema são estruturados à priori e, portanto, os seus actos de diálogo são mais fáceis de prever do que os dos segmentos do utilizador. Para além disso, um sistema de diálogo está ciente dos seus próprios actos de diálogo e tem apenas de prever os dos segmentos dos seus interlocutores. Logo, nesse cenário, apenas o desempenho nos segmentos do utilizador é relevante. Como esperado, o desempenho foi mais elevado nos segmentos do sistema em todos os níveis. No entanto, nos segmentos do utilizador, a taxa de acerto no Nível 1 e a taxa de correspondência exacta nos restantes níveis ainda ficaram acima de 90%. Para além disso, é importante referir que, como o Nível 3 é altamente relacionado com palavras específicas, a diferença de desempenho não é tão pronunciada nesse nível.

Por fim, ao combinar hierarquicamente os melhores classificadores para cada nível, obtivemos, em média, uma taxa de correspondência exacta de 94,28% na combinação dos Níveis 1 e 2 e 92,34% na combinação dos três níveis. Esses resultados são já em linha com ou superiores aos obtidos em estudos anteriores sobre o reconhecimento de actos de diálogo no corpus DIHANA. No entanto, esses estudos consideraram uma versão simplificada do problema, reduzindo-o a um problema de classificação de etiqueta única, em que a etiqueta de um segmento consiste na concatenação das etiquetas dos três níveis. Uma vez que esta abordagem considera apenas as combinações de etiquetas presentes no corpus, o número de etiquetas possíveis é reduzido em comparação com a nossa abordagem, que aborda a previsão de etiquetas dos Níveis 2 e 3 como problemas de classificação multi-etiqueta. Ao abordar o problema de forma comparável à desses estudos, os valores anteriores aumentam para 96,24% e 93,98%, respectivamente.

Como trabalho futuro, seria interessante avaliar se as conclusões tiradas deste estudo sobre dados em espanhóis e, anteriormente, sobre dados em inglês, se mantêm para dados em outras línguas com tipologia morfológica diferente. Em termos de reconhecimento de actos de diálogo multi-etiqueta, seria interessante explorar o uso de outras funções de custo ao treinar a rede, especialmente uma baseada na medida-F, que não é tão influenciada pelo número reduzido de classes positivas por segmento como a função de custo de Hamming. Para além disso, é importante avaliar se as abordagens de representação do segmento baseadas em *tokenização* ao nível do carácter

são capazes de capturar informação adicional para prever as etiquetas específicas da tarefa. Também seria interessante explorar meios para realizar a classificação hierárquica dos múltiplos níveis usando uma única rede em vez de três classificadores independentes. Por fim, é importante avaliar a deterioração do desempenho num cenário real. Ou seja, um em que o sistema de diálogo não é simulado e, portanto, tem de lidar com problemas relacionados com o Reconhecimento Automático de Fala (ASR) e usar etiquetas previstas automaticamente como informação de contexto.

Agradecimentos

Este estudo foi financiado por fundos nacionais, através da Fundação para a Ciência e a Tecnologia (FCT), com a referência UID/CEC/50021/2019, e pela Universidade de Lisboa.

Referências

Abadi, Martín et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/.

Alcácer, N., J. M. Benedí, F. Blat, R. Granell, C. D. Martínez & F. Torres. 2005. Acquisition and Labelling of a Spontaneous Speech Dialogue Corpus. Em *SPECOM*, 583–586.

Alexandersson, Jan, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Michael Kipp, Stephan Koch, Elisabeth Maier, Norbert Reithinger, Birte Schmitz & Melanie Siegel. 1998. Dialogue Acts in VERBMOBIL-2 Second Edition. Relatório técnico. DFKI.

Anderson, Anne H., Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller et al. 1991. The HCRC Map Task Corpus. *Language and Speech* 34(4). 351–366.

Ang, Jeremy, Yang Liu & Elizabeth Shriberg. 2005. Automatic Dialog Act Segmentation and Classification in Multiparty Meetings. Em *ICASSP*, vol. 1, 1061–1064.

Benedí, José-Miguel, Eduardo Lleida, Amparo Varona, María-José Castro, Isabel Galiano, Raquel Justo, Iñigo López de Letona & Antonio Miguel. 2006. Design and Acquisition of a Telephone Spontaneous Speech Dialogue Corpus in Spanish: DIHANA. Em *LREC*, 1636–1639.

Bunt, Harry, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis & David R. Traum. 2012. ISO 24617-2: A Semantically-Based Standard for Dialogue Annotation. Em *LREC*, 430–437.

Bunt, Harry, Volha Petukhova, Andrei Malchanau, Kars Wijnhoven & Alex Fang. 2016. The DialogBank. Em *LREC*, 3151–3158.

Cardellino, Cristian. 2016. Spanish Billion Word Corpus and Embeddings. http://crscardellino.me/SBWCE/.

Carletta, Jean, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma & Pierre Wellner. 2005. The AMI Meeting Corpus: A Pre-Announcement. Em *MLMI*, 28–39.

Carroll, John M. & Michael K. Tanenhaus. 1978. Functional Clauses and Sentence Segmentation. *Journal of Speech, Language, and Hearing Research* 21(4). 793–808.

Chollet, François et al. 2015. Keras: The Python Deep Learning Library. https://keras.io/.

Conneau, Alexis, Holger Schwenk, Loïc Barrault & Yann Lecun. 2017. Very Deep Convolutional Networks for Text Classification. Em *EACL*, vol. 1, 1107–1116.

Coria, Sergio R. & Luis A. Pineda. 2005. Predicting Obligation Dialogue Acts from Prosodic and Speaker Information. *Research in Computing Science* 14. 137–148.

Coria, Sergio R. & Luis A. Pineda. 2006. Predicting Dialogue Acts from Prosodic Information. Em *CICLing*, 355–365.

Coria, Sergio R. & Luis A. Pineda. 2009. An Analysis of Prosodic Information for the Recognition of Dialogue Acts in a Multimodal Corpus in Mexican Spanish. *Computer Speech & Language* 23(3). 277–310.

Costantini, Erica, Susanne Burger & Fabio Pianesi. 2002. NESPOLE!'s Multilingual and Multimodal Corpus. Em *LREC*, 165–170.

Di Eugenio, Barbara, Zhuli Xie & Riccardo Serafin. 2010. Dialogue Act Classification, Higher Order Dialogue Structure, and Instance-Based Learning. *Dialogue and Discourse* 1(2). 81–104.

Díez, Jorge, Oscar Luaces, Juan José del Coz & Antonio Bahamonde. 2015. Optimizing Different Loss Functions in Multilabel Classifications. *Progress in Artificial Intelligence* 3(2). 107–118.

Gambäck, Björn, Fredrik Olsson & Oscar Täckström. 2011. Active Learning for Dialogue Act Classification. Em *INTERSPEECH*, 1329–1332.

Jekat, Susanne, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast & J. Joachim Quantz. 1995. Dialogue Acts in VERBMOBIL. Relatório técnico. DFKI.

Ji, Yangfeng, Gholamreza Haffari & Jacob Eisenstein. 2016. A Latent Variable Recurrent Neural Network for Discourse Relation Language Models. Em *NAACL-HLT*, 332–342.

Jurafsky, Dan, Elizabeth Shriberg & Debra Biasca. 1997. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual. Relatório Técnico. Draft 13 University of Colorado, Institute of Cognitive Science.

Kalchbrenner, Nal & Phil Blunsom. 2013. Recurrent Convolutional Neural Networks for Discourse Compositionality. Em *Workshop on Continuous Vector Space Models and their Compositionality*, 119–126.

Kay, Martin, Peter Norvig & Mark Gawron. 1992. *VERBMOBIL: A Translation System for Face-to-Face Dialog*. University of Chicago Press.

Khanpour, Hamed, Nishitha Guntakandla & Rodney Nielsen. 2016. Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network. Em *COLING*, 2012–2021.

Kim, Seokhwan, Luis Fernando D'Haro, Rafael E. Banchs, Jason Williams & Matthew Henderson. 2016. The Fourth Dialog State Tracking Challenge. Em *International Workshop on Spoken Dialog Systems*, .

Kim, Yoon. 2014. Convolutional Neural Networks for Sentence Classification. Em *EMNLP*, 1746–1751.

Kingma, Diederik P. & Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. Em *ICLR*, http://arxiv.org/abs/1412.6980.

Král, Pavel & Christophe Cerisara. 2010. Dialogue Act Recognition Approaches. *Computing and Informatics* 29(2). 227–250.

Lee, Ji Young & Franck Dernoncourt. 2016. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. Em *NAACL-HLT*, 515–520.

Levin, Lori, Ann Thymé-Gobbel, Alon Lavie, Klaus Ries & Klaus Zechner. 1998. A Discourse Coding Scheme for Conversational Spanish. Em *ICSLP*, paper 1000.

Levin, Lori S., Klaus Ries, Ann Thymé-Gobbel & Alon Lavie. 1999. Tagging Of Speech Acts And Dialogue Games In Spanish Call Home. Em *Workshop On Towards Standards And Tools For Discourse Tagging*, 42–47.

Liu, Yang, Kun Han, Zhao Tan & Yun Lei. 2017. Using Context Information for Dialog Act Classification in DNN Framework. Em *EMNLP*, 2160–2168.

Margolis, Anna, Karen Livescu & Mari Ostendorf. 2010. Domain Adaptation with Unlabeled Data for Dialog Act Tagging. Em *Workshop on Domain Adaptation for Natural Language Processing* DANLP 2010, 45–52.

Martínez-Hinarejos, Carlos D., José-Miguel Benedí & Ramón Granell. 2008. Statistical Framework for a Spanish Spoken Dialogue Corpus. *Speech Communication* 50(11–12). 992–1008.

Martínez-Hinarejos, Carlos D., José-Miguel Benedí & Vicent Tamarit. 2015. Unsegmented Dialogue Act Annotation and Decoding with N-Gram Transducers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23(1). 198–211.

Martínez-Hinarejos, Carlos D., Emilio Sanchis, Fernando García-Granada & Pablo Aibar. 2002. A Labelling Proposal to Annotate Dialogues. Em *LREC*, vol. 5, 1566–1582.

Mezza, Stefano, Alessandra Cervone, Evgeny A. Stepanov, Giuliano Tortoreto & Giuseppe Riccardi. 2018. ISO-Standard Domain-Independent Dialogue Act Tagging for Conversational Agents. Em *COLING*, 3539–3551.

Mikolov, Tomas, Martin Karafiát, Lukás Burget, Jan Cernocký & Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. Em *INTERSPEECH*, 1045–1048.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado & Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. Em *NIPS*, 3111–3119.

Pennington, Jeffrey, Richard Socher & Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. Em *EMNLP*, 1532–1543.

Petukhova, Volha, Martin Gropp, Dietrich Klakow, Anna Schmidt, Gregor Eigner, Mario Topf, Stefan Srb, Petr Motlicek, Blaise Potard, John Dines, Olivier Deroo, Ronny Egeler, Uwe Meinz & Steffen Liersch. 2014. The DBOX Corpus Collection of Spoken Human-Human and Human-Machine Dialogues. Em *LREC*, 252–258.

Ribeiro, Eugénio, Ricardo Ribeiro & David Martins de Matos. 2015. The Influence of Context on Dialogue Act Recognition. *CoRR* abs/1506.00839. http://arxiv.org/abs/1506.00839.

Ribeiro, Eugénio, Ricardo Ribeiro & David Martins de Matos. 2016. Mapping the Dialog Act Annotations of the LEGO Corpus into the Communicative Functions of ISO 24617-2. *CoRR* abs/1612.01404. http://arxiv.org/abs/1612.01404.

Ribeiro, Eugénio, Ricardo Ribeiro & David Martins de Matos. 2018. A Study on Dialog Act Recognition using Character-Level Tokenization. Em *AIMSA*, 93–103.

Ries, Klaus. 1999. HMM and Neural Network Based Speech Act Detection. Em *ICASSP*, vol. 1, 497–500.

Schmitt, Alexander, Stefan Ultes & Wolfgang Minker. 2012. A Parameterized and Annotated Spoken Dialog Corpus of the CMU Let's Go Bus Information System. Em *LREC*, 3369–3373.

Searle, John R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Serafin, Riccardo & Barbara Di Eugenio. 2004. FLSA: Extending Latent Semantic Analysis with Features for Dialogue Act Classification. Em *ACL*, 692–699.

Serafin, Riccardo, Barbara Di Eugenio & Michael Glass. 2003. Latent Semantic Analysis for Dialogue Act Classification. Em *NAACL-HLT*, vol. 2, 94–96.

Shriberg, Elizabeth, Raj Dhillon, Sonali Bhagat, Jeremy Ang & Hannah Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. Em *SIGDIAL*, 97–100.

Sorower, Mohammad S. 2010. A Literature Survey on Algorithms for Multi-Label Learning. Relatório técnico. Oregon State University.

Stolcke, Andreas, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin & Marie Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics* 26(3). 339–373.

Tamarit, Vicent & Carlos D. Martínez-Hinarejos. 2008. Dialog Act Labeling in the DIHANA Corpus using Prosody Information. Em *V Jornadas en Tecnología del Habla*, 183–186.

Tran, Quan Hung, Ingrid Zukerman & Gholamreza Haffari. 2017a. A Generative Attentional Neural Network Model for Dialogue Act Classification. Em *ACL*, vol. 2, 524–529.

Tran, Quan Hung, Ingrid Zukerman & Gholamreza Haffari. 2017b. A Hierarchical Neural Model for Learning Sequences of Dialogue Acts. Em *EACL*, vol. 1, 428–437.

Tran, Quan Hung, Ingrid Zukerman & Gholamreza Haffari. 2017c. Preserving Distributional Information in Dialogue Act Classification. Em *EMNLP*, 2151–2156.

Villaseñor, Luis, Antonio Massé & Luis A. Pineda. 2001. The DIME Corpus. Em *Encuentro Internacional de Ciencias de la Computación*, vol. 2, 1–10.