information

MDPI

Article

A Multilingual and Multidomain Study on Dialog Act Recognition Using Character-Level Tokenization †

Eugénio Ribeiro [1,2,*], Ricardo Ribeiro [1,3] and David Martins de Matos [1,2]

[1] L^2F—Spoken Language Systems Laboratory—INESC-ID, 1000-029 Lisboa, Portugal; ricardo.ribeiro@inesc-id.pt (R.R.); david.matos@inesc-id.pt (D.M.d.M.)
[2] Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisboa, Portugal
[3] Instituto Universitário de Lisboa (ISCTE-IUL), 1649-026 Lisboa, Portugal
* Correspondence: eugenio.ribeiro@l2f.inesc-id.pt

† This paper is an extended version of our paper published in The 18th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA 2018).

Received: 21 January 2019; Accepted: 26 February 2019; Published: 3 March 2019

Abstract: Automatic dialog act recognition is an important step for dialog systems since it reveals the intention behind the words uttered by its conversational partners. Although most approaches on the task use word-level tokenization, there is information at the sub-word level that is related to the function of the words and, consequently, their intention. Thus, in this study, we explored the use of character-level tokenization to capture that information. We explored the use of multiple character windows of different sizes to capture morphological aspects, such as affixes and lemmas, as well as inter-word information. Furthermore, we assessed the importance of punctuation and capitalization for the task. To broaden the conclusions of our study, we performed experiments on dialogs in three languages—English, Spanish, and German—which have different morphological characteristics. Furthermore, the dialogs cover multiple domains and are annotated with both domain-dependent and domain-independent dialog act labels. The achieved results not only show that the character-level approach leads to similar or better performance than the state-of-the-art word-level approaches on the task, but also that both approaches are able to capture complementary information. Thus, the best results are achieved by combining tokenization at both levels.

Keywords: dialog act recognition; character-level; multilinguality; multidomain

1. Introduction

Dialog act recognition is an important task in the context of a dialog system, since dialog acts are the minimal units of linguistic communication that reveal the intention behind the uttered words [1]. Identifying the intention behind the utterances of its conversational partners allows a dialog system to apply specialized interpretation strategies, accordingly. Thus, automatic dialog act recognition is a task that has been widely explored over the years on multiple corpora and using multiple classical machine learning approaches [2]. However, recently, most approaches on the task focus on applying different Deep Neural Network (DNN) architectures to generate segment representations from word embeddings and combine them with context information from the surrounding segments [3–6]. All of these approaches look at the segment at the word level. That is, they consider that a segment is a sequence of words and that its intention is revealed by the combination of those words. However, there are also cues for intention at the sub-word level. These cues are mostly related to the morphology of words. For instance, there are cases, such as adverbs of manner and negatives, in which the function, and hence the intention, of a word is related to its affixes. On the other hand, there are cases in which considering multiple forms of the same lexeme independently does not provide additional information concerning intention and the lemma suffices. This information is provided by different groups of characters that constitute relevant morphological aspects. Thus, it is hard to capture using word-level approaches.

We aimed at capturing sub-word information and, consequently, improving the state-of-the-art on dialog act recognition by performing character-level tokenization and exploring the use of multiple context windows surrounding each token to capture different morphological aspects. Although character-level approaches are typically used for word-level classification tasks, such as Part-of-Speech (POS) tagging [7], their use for dialog act recognition is supported by the interesting results achieved on other short-text classification tasks, such as language identification [8] and review rating [9]. In addition to the aspects concerning morphological information, using character-level tokenization allows assessing the importance of aspects such as capitalization and punctuation for the task. Furthermore, we assessed whether the character-level approach is able to capture all the relevant word-level information or if there are some aspects that can only be captured using a word-level approach. That is, we assessed whether the word- and character-level approaches are complementary and can be combined to improve the performance on the task.

In the conference paper that this article extends [10], we performed experiments on two corpora, Switchboard Dialog Act Corpus (SwDA) [11] and DIHANA [12], which vary in terms of domain, the nature of the participants, and language—English and Spanish, respectively. However, in both cases, we focused on assessing the performance of character-level approaches to predict generic domain-independent dialog act labels. Here, we extended that study by assessing the performance when predicting the domain-dependent dialog act labels of the LEGO corpus [13] (an annotated subset of the Let's Go Corpus) and the Levels 2 and 3 of the dialog act annotations of the DIHANA corpus, which are both domain-dependent and multilabel. Additionally, we assessed the performance on the German dialogs of the VERBMOBIL corpus [14]. The latter is interesting since the German language features morphological aspects that are not predominant in English or Spanish.

In the remainder of the article, we start by providing an overview of previous approaches on dialog act recognition, in Section 2. Then, in Section 3, we discuss why using character-level tokenization is relevant for the task and define the aspects that we want to explore. Section 4 describes our experimental setup, including the used datasets in Section 4.1, our classification approach in Section 4.2, and the evaluation method and word-level baselines in Section 4.3. The results of our experiments are presented and discussed in Section 5. Finally, Section 6 states the most important conclusions of this study and provides pointers for future work.

2. Related Work

Automatic dialog act recognition is a task that has been widely explored over the years, using multiple classical machine learning approaches, from Hidden Markov Models (HMMs) [15] to Support Vector Machines (SVMs) [16,17]. The article by Král and Cerisara [2] provides an interesting overview on many of those approaches on the task. However, recently, similar to many other Natural Language Processing (NLP) tasks [18,19], most studies on dialog act recognition take advantage of different Neural Network (NN) architectures.

To our knowledge, the first of those studies was that by Kalchbrenner and Blunsom [3]. The described approach uses a Convolutional Neural Network (CNN)-based approach to generate segment representations from randomly initialized word embeddings. Then, it uses a Recurrent Neural Network (RNN)-based discourse model that combines the sequence of segment representations with speaker information and outputs the corresponding sequence of dialog acts. By limiting the discourse model to consider information from the two preceding segments only, this approach achieved 73.9% accuracy on the SwDA corpus.

Lee and Dernoncourt [4] compared the performance of a Long Short-Term Memory (LSTM) unit against that of a CNN to generate segment representations from pre-trained embeddings of its words. To generate the corresponding dialog act classifications, the segment representations were then fed to a two-layer feed-forward network, in which the first layer normalizes the representations and the second selects the class with highest probability. In their experiments, the CNN-based approach consistently led to similar or better results than the LSTM-based one. The architecture was also used to provide context information from up to two preceding segments at two levels. The first level refers to the concatenation of the representations of the preceding segments with that of the current segment before providing it to the feed-forward network. The second refers to the concatenation of the normalized representations before providing them to the output layer. This approach achieved 65.8%, 84.6%, and 71.4% accuracy on the Dialog State Tracking Challenge 4 (DSTC4) [20], ICSI Meeting Recorder Dialog Act Corpus (MRDA) [21], and SwDA corpora, respectively. However, the influence of context information varied across corpora.

Ji et al. [5] explored the combination of positive aspects of NN architectures and probabilistic graphical models. They used a Discourse Relation Language Model (DRLM) that combined a Recurrent Neural Network Language Model (RNNLM) [22] to model the sequence of words in the dialog with a latent variable model over shallow discourse structure to model the relations between adjacent segments which, in this context, are the dialog acts. This way, the model can perform word prediction using discriminatively-trained vector representations while maintaining a probabilistic representation of a targeted linguistic element, such as the dialog act. To function as a dialog act classifier, the model was trained to maximize the conditional probability of a sequence of dialog acts given a sequence of segments, achieving 77.0% accuracy on the SwDA corpus.

The previous studies explored the use of a single recurrent or convolutional layer to generate the segment representation from those of its words. However, the top performing approaches use multiple of those layers. On the one hand, Khanpour et al. [23] achieved their best results using a segment representation generated by concatenating the outputs of a stack of 10 LSTM units at the last time step. This way, the model is able to capture long distance relations between tokens. On the convolutional side, Liu et al. [6] generated the segment representation by combining the outputs of three parallel CNNs with different context window sizes to capture different functional patterns. In both cases, pre-trained word embeddings were used as input to the network. Overall, from the reported results, it is not possible to state which is the top performing segment representation approach since the evaluation was performed on different subsets of the SwDA corpus. However, Khanpour et al. [23] reported 73.9% accuracy on the validation set and 80.1% on the test set, while Liu et al. [6] reported 74.5% and 76.9% accuracy on the two sets used to evaluate their experiments. Additionally, Khanpour et al. [23] reported 86.8% accuracy on the MRDA corpus.

Liu et al. [6] also explored the use of context information concerning speaker changes and from the surrounding segments. The first was provided as a flag and concatenated to the segment representation. Concerning the latter, they explored the use of discourse models, as well as of approaches that concatenated the context information directly to the segment representation. The discourse models transform the model into a hierarchical one by generating a sequence of dialog act classifications from the sequence of segment representations. Thus, when predicting the classification of a segment, the surrounding ones are also taken into account. However, when the discourse model is based on a CNN or a bidirectional LSTM unit, it considers information from future segments, which is not available to a dialog system. However, even when relying on future information, the approaches based on discourse models performed worse than those that concatenated the context information directly to the segment representation. In this sense, providing that information in the form of the classification of the surrounding segments led to better results than using their words, even when those classifications were obtained automatically. This conclusion is in line with what we had shown in our previous study using SVMs [17]. Furthermore, both studies have shown that, as expected, the first preceding segment is the most important and that the influence decays with the distance. Using the setup with gold standard labels from three preceding segments, the results on the two sets used to evaluate the approach improved to 79.6% and 81.8%, respectively.

Finally, considering the focus of this study, it is important to make some remarks concerning tokenization and token representation. In all the previously described studies, tokenization was performed at the word level. Furthermore, with the exception of the first study [3], which used randomly initialized embeddings, the representation of those words was given by pre-trained embeddings. Khanpour et al. [23] compared the performance when using Word2Vec [24] and Global Vectors for Word Representation (GloVe) [25] embeddings trained on multiple corpora. Although both embedding approaches capture information concerning words that commonly appear together, the best results were achieved using Word2Vec embeddings. In terms of dimensionality, that study compared embedding spaces with 75, 150, and 300 dimensions. The best results were achieved when using 150-dimensional embeddings. However, 200-dimensional embeddings were used in other studies [4,6], which was not one of the compared values.

3. The Character Level

It is interesting to explore character-level tokenization because it allows us to capture morphological information that is at the sub-word level and, thus, cannot be directly captured using word-level tokenization. Considering the task at hand, that information is relevant since it may provide cues for identifying the intention behind the words. When someone selects a set of words to form a segment that transmits a certain intention, each of those words is typically selected because it has a function that contributes to that transmission. In this sense, affixes are tightly related to word function, especially in fusional languages. Thus, the presence of certain affixes is a cue for intention, independently of the lemma. However, there are also cases, such as when affixes are used for subject–verb agreement, in which the cue for intention is in the lemmas and, thus, considering multiple forms of the same lexeme does not provide additional information.

Information concerning lemmas and affixes cannot be captured from single independent characters. Thus, it is necessary to consider the context surrounding each token and look at groups of characters. The size of the context window plays an important part in what information can be captured. For instance, English affixes are typically short (e.g., *un-*, *-ly*, and *a-*), but in other languages, such as Spanish, there are longer commonly used affixes (e.g., *-mente*, which is used to transform adjectives into adverbs in the same manner as *-ly* in English). Furthermore, to capture the lemmas of long words, the agglutinative aspects of German, and even inter-word relations, wider context window sizes must be considered. However, using wide context windows impairs the ability to capture information from short groups of characters, as additional irrelevant characters are considered. This suggests that, to capture all the relevant information, multiple context windows should be used.

Using character-level tokenization also allows us to consider punctuation, which is able to provide both direct and indirect cues for dialog act recognition. For instance, an interrogation mark provides a direct cue that the intention is related to knowledge seeking. On the other hand, commas structure the segment, indirectly contributing to the transmission of an intention.

Additionally, character-level tokenization allows us to consider capitalization information. However, in the beginning of a segment, capitalization only signals that beginning and, thus, considering it only introduces entropy. In both English and Spanish, capitalization in the middle of a segment is typically only used to distinguish proper nouns, which are not related to intention. In German, all nouns are capitalized, which simplifies their distinction from words of other classes, such as adjectives. However, while some nouns may be related to intention, others are not. Thus, capitalization information is not expected to contribute to the task.

Finally, previous studies have shown that word-level information is relevant for the task and that word-level approaches are able to identify intention with acceptable performance. Although we expect character-level approaches to be able to capture most of the information that is captured at the word level, exploring the character level highly increases the number of tokens in a segment, which introduces a large amount of entropy. Thus, it is possible that some specific aspects can only be captured at the word level. In this case, it is important to assess whether both approaches are complementary.

4. Experimental Setup

To assess the validity of the hypotheses proposed in the previous section, we performed experiments on different corpora and compared the performance when using word- and character-level tokenization. The used datasets, the classification and evaluation approaches, and the word-level baselines are described below.

4.1. Datasets

In the conference paper extended by this article [10], we performed experiments on two corpora, SwDA [11] and DIHANA [12]. These datasets allowed us to assess the performance of character-level approaches when predicting generic domain-independent dialog act labels in both English and Spanish. In this article, we extended our study by exploring an additional language, German. To do so, we relied on the VERBMOBIL corpus [14]. Additionally, we assessed the performance of character-level approaches when predicting the domain-specific dialog act labels of the LEGO corpus [13] and the lower levels of the dialog act annotation of the DIHANA corpus. Each of these corpora and their dialog act annotations are described below.

4.1.1. Switchboard Dialog Act Corpus

The Switchboard [26] corpus consists of about 2400 telephone conversations among 543 American English speakers. Each pair of speakers was automatically attributed a topic for discussion, from 70 different ones. Furthermore, speaker pairing and topic attribution were constrained so that no two speakers would be paired with each other more than once and no one spoke more than once on a given topic. The Switchboard Dialog Act Corpus (SwDA) [11] is a subset of this corpus, consisting of 1155 manually transcribed conversations, containing 223,606 segments.

The corpus was annotated for dialog acts using the SWBD-DAMSL tag set, which was structured so that the annotators were able to label the conversations from transcriptions alone. It contains over 200 unique tags. However, to obtain a higher inter-annotator agreement and higher example frequencies per class, a less fine-grained set of 44 tags was devised. Jurafsky et al. [11] reported an average pairwise Kappa [27] of 0.80, while Stolcke et al. [15] referred to an inter-annotator agreement of 84%, which is the average pairwise percent agreement. As shown in Table 1, the class distribution is highly unbalanced, with the three most frequent classes—*Statement-opinion*, *Acknowledgement*, and *Statement-non-opinion*—covering 68% of the corpus. The tag set is typically further reduced to 42 categories [15], by merging the *Abandoned* and *Uninterpretable* categories and by merging the segments labeled as *Segment* with the previous one by the same speaker. Although there are other variations of the tag set [16,28,29], we used the 42-label version in our experiments, since by analyzing the data we came to the conclusion that it is the most appropriate.

We selected this corpus for our experiments because it is the most explored for the dialog act recognition task, since it contains a large amount of annotated data, which can lead to solid results. Furthermore, since its tag set is domain-independent and the domain of the dialog varies, the probability of drawing conclusions that depend on the domain of the corpus is reduced.

Table 1. Label distribution in the Switchboard Dialog Act Corpus (SwDA) [11].

Label	Count	%	Label	Count	%
Statement-non-opinion	72,824	36	Collaborative Completion	699	0.4
Acknowledgement	37,096	19	Repeat-Phrase	660	0.3
Statement-opinion	25,197	13	Open-Question	632	0.3
Agreement	10,820	5	Rhetorical-Question	557	0.3
Abandoned	10,569	5	Hold	540	0.2
Appreciation	4663	2	Reject	338	0.2
Yes-No-Question	4624	2	Negative Non-no Answer	292	0.1
Non-verbal	3548	2	Non-understanding	288	0.1
Yes Answer	2934	1	Other Answer	279	0.1
Conventional Closing	2486	1	Conventional Opening	220	0.1
Uninterpretable	2158	1	Or-Clause	207	0.1
Wh-Question	1911	1	Dispreferred Answers	205	0.1
No Answer	1340	1	3rd-party-talk	115	0.1
Response Acknowledgement	1277	1	Offers/Options	109	0.1
Hedge	1182	1	Self-talk	102	0.1
Declarative Yes-No-Question	1174	1	Downplayer	100	0.1
Other	1074	1	Maybe	98	<0.1
Backchannel-Question	1019	1	Tag-Question	93	<0.1
Quotation	934	0.5	Declarative Wh-Question	80	<0.1
Summarization	919	0.5	Apology	76	<0.1
Affirmative Non-yes Answer	836	0.4	Thanking	67	<0.1
Action Directive	719	0.4			

4.1.2. LEGO

The LEGO corpus [13] is an annotated subset of 347 calls from the Carnegie Mellon University (CMU)’s Let’s Go Bus Information System [30] recorded during 2006. It features 14,186 utterances—9083 system utterances and 5103 user utterances. Since system utterances are generated through slot filling of fixed templates, they have no errors and contain casing and punctuation information. In contrast, the transcriptions of user utterances were obtained using an Automatic Speech Recognition (ASR) system and, thus, contain no casing or punctuation information. Furthermore, the recognition was not always correct. However, a concrete value for the Word Error Rate (WER) is not revealed.

In terms of dialog acts, the LEGO corpus is annotated with two distinct and domain-dependent tag sets for system and user turns. The set for system turns contains 28 labels, while the set for user turns contains 22 labels. Since the two sets are distinct and the system is aware of the dialog acts of its own sentences, it makes no sense to predict them automatically. Thus, we focused on the user segments and only relied on system segments to provide context information. The distribution of the labels across user segments is shown in Table 2. The three most common labels are *Place Information*, *Unqualified/Unrecognized*, and *Reject*, with the last two covering 29% of the segments. This reveals a high number of communication problems between the user and the system.

Table 2. Label distribution in the user segments of the LEGO corpus.

Label	Count	%	Label	Count	%
Place Information	879	17	Reject Departure	135	3
Unqualified/Unrecognized	783	15	New Query	98	2
Reject	746	14	Reject Time	95	2
Line Information	440	8	Request Previous Bus	75	1
Time Information	391	8	Reject Bus	69	1
Confirm Departure	291	6	Reject Destination	53	1
Confirm Destination	246	5	Request Help	52	1
Confirm Time	225	4	Goodbye	29	0.6
Confirm	214	4	Request Schedule	18	0.4
Confirm Bus	179	3	Polite	8	0.2
Request Next Bus	159	3	Inform	3	0.1

Although the LEGO corpus has been used in many research tasks related to dialog and interaction with Interactive Voice Response (IVR) systems, its dialog act annotations have been neglected. In fact, to our knowledge, they have only been used in the context of the SpeDial project [31] (http://www.spedial.eu). This is probably due to the domain dependence of the labels, which would not be useful in any other domain. Furthermore, since the labels are so specific, even a system dealing with the same domain would only be able to benefit from them if the dialog had the same characteristics as the ones from the LEGO corpus. However, we decided to use this corpus in our experiments, since it allowed us to explore the performance of character-based approaches when predicting domain-dependent dialog acts.

4.1.3. DIHANA

The DIHANA corpus [12] consists of 900 dialogs between 225 human speakers and a Wizard of Oz (WoZ) telephonic train information system. There are 6280 user turns and 9133 system turns, with a vocabulary size of 823 words and a total of 48,243 words. The turns were manually transcribed, segmented, and annotated with dialog acts [32]. The total number of annotated segments is 23,547, with 9715 corresponding to user segments and 13,832 to system segments.

Contrarily to what happens in the LEGO corpus, the dialog act annotations are common to both user and system segments. Furthermore, they are hierarchically decomposed in three levels [33]. Level 1 represents the generic intention of the segment, independently of task details, while the remaining represent task-specific information. Level 1 has 11 labels, distributed according to Table 3. In that table, we can see that two of the labels are exclusive to user segments—*Acceptance* and *Rejection*—and four to system segments—*Opening*, *Waiting*, *New Consult*, and *Confirmation*. Furthermore, the most common label, *Question*, covers 27% of the segments.

Table 3. Distribution of the domain-independent labels of Level 1 in the DIHANA corpus. The labels shown in this table are translations of the original labels in Spanish.

Label	User	System	Total	%	Label	User	System	Total	%
Question	5474	864	6338	27	Acceptance	990	0	990	4
Answer	1839	2446	4285	18	Opening	0	900	900	4
Confirmation	0	3629	3629	15	Not Understood	4	653	657	3
New Consult	0	2474	2474	11	Rejection	340	0	340	1
Waiting	0	1948	1948	8	Undefined	141	18	159	1
Closing	927	900	1827	8					

Although they share most labels, the two task-specific levels of the hierarchy focus on different information. While Level 2 is related to the kind of information that is implicitly focused in the segment, Level 3 is related to the kind of information that is explicitly referred to in the segment. For instance, consider the segment *"I'm looking for trains departing from Bilbao to Corunna on Monday, February 16, 2004."* Since it reveals the intention of finding a train schedule, it has *Departure Time* as a Level 2 label. However, since that departure time is not explicitly referred in the segment, that label is not part of its Level 3 labels. On the other hand, the segment explicitly refers a departure place, a destination, and a date. Thus, it has the corresponding Level 3 labels—*Origin*, *Destination*, and *Day*.

The distributions of the labels of both levels in the corpus are shown in Table 4. We can see that there are 10 common labels and three additional ones on Level 3—*Order Number*, *Number of Trains*, and *Trip Type*. Furthermore, both levels have the *Nil* label, which represents the absence of label in that level. In this sense, we can see that only 63% of the segments have Level 2 labels, and that the percentage is even lower, 52%, when considering Level 3 labels. This is mainly because the segments labeled as *Opening*, *Closing*, *Undefined*, *Not Understood*, *Waiting*, and *New Consult* on the first level cannot have labels on the remaining levels. This happens since those labels refer to discourse structuring or problems in the dialog. Finally, it is important to note that, while each segment may only have one Level 1 label, it may have multiple Level 2 and Level 3 labels.

Table 4. Distribution of the domain-dependent labels of Levels 2 and 3 in the DIHANA corpus. The labels shown in this table are translations of the original labels in Spanish.

Level 2					**Level 3**				
Label	**User**	**System**	**Total**	**%**	**Label**	**User**	**System**	**Total**	**%**
Nil	1923	6893	8816	37	Nil	2954	8317	11,271	48
Departure Time	3309	3523	7432	32	Destination	1631	2079	3710	16
Fare	2071	1267	3338	14	Day	1881	1778	3659	16
Day	1026	923	1949	8	Origin	896	2085	2981	13
Origin	477	480	957	4	Departure Time	692	1633	2325	10
Destination	452	400	852	4	Number of Trains	0	1863	1863	8
Train Type	317	226	543	2	Train Type	544	1253	1797	8
Arrival Time	90	88	178	1	Order Number	84	950	1034	4
Duration	14	15	29	0.1	Ticket Class	129	766	895	4
Ticket Class	15	12	27	0.1	Fare	47	731	778	3
Service	3	5	8	<0.1	Arrival Time	199	490	689	3
					Trip Type	643	0	643	3
					Service	15	4	19	0.1
					Duration	0	14	14	0.1

We decided to use the DIHANA corpus in our experiments, since it allowed us to study the identification of both domain-independent and domain-dependent dialog acts in a language other than English. Furthermore, it poses multilabel classification problems at the lower levels, which are not common and are interesting to explore.

4.1.4. VERBMOBIL

The VERBMOBIL project [14] aimed at developing a portable translation device that could be used in meetings involving speakers of different languages. During its first phase, multiple dialogs in an appointment scheduling scenario were recorded. The segments in those dialogs were annotated with dialog act information using a taxonomy featuring 18 domain-independent labels, some of which could then be further specified using domain information, leading to a set of 42 domain-dependent labels [34]. The scenario was later extended in the second phase of the project to dialogs in the travel planning domain, of which appointment scheduling is only a part. Thus, the dialog act annotations had to be updated in order to use a common set of labels for dialogs of both phases of the project. The updated label set consists of 33 domain-independent labels, which can be attributed to segments using a decision tree, stopping when there is not enough information to go deeper in the tree [35].

Although the corpus also features dialogs in English and Japanese, most are in German. In this sense, there are 735 dialogs in German, featuring 39,148 segments, annotated with the updated version of the label set. This is the corpus we considered for our experiments. To avoid dealing with a hierarchical problem, we collapsed the 33 labels into the 17 at the top level of the tree. Table 5 shows the distribution of these labels in the corpus. We can see that the three most common labels—*Feedback*, *Inform*, and *Suggest*—cover 66% of the corpus and are highly related to the planning/scheduling nature of the dialogs.

We decided to include the VERBMOBIL corpus in our study since it provides dialogs in German, which differs from English and Spanish in terms of morphology. More specifically, although it is a fusional language, agglutinative aspects are much more predominant in German than in the other two languages.

Table 5. Label distribution in the VERBMOBIL corpus.

Label	Count	%	Label	Count	%
Feedback	10,002	26	Not Classifiable	918	2
Inform	7819	20	Introduce	713	2
Suggest	7815	20	Thank	452	1
Request	3314	8	Close	420	1
Bye	1613	4	Politeness Formula	363	1
Init	1458	4	Commit	254	0.6
Greet	1408	4	Defer	176	0.4
Deliberate	1319	3	Offer	37	0.1
Backchannel	1067	3			

4.2. Classification Approach

As a classification approach, we adapted the CNN-based word-level approach by Liu et al. [6] to use characters instead of words as tokens. We opted for this approach since its multiple parallel CNNs are able to capture information from sets of characters of different sizes, which is important to capture the different morphological aspects referred to in Section 3.

The generic architecture of the network used in both the word-level baselines and our experiments at the character-level is summarized in Figure 1. Its input is the sequence of embeddings of the tokens in a segment. The embedding representation at the word level is part of the definition of the word-level baselines. Thus, it is described in the next section. At the character level, instead of using a one-hot encoding for each character, we use the number of dimensions required for a one-hot encoding, but allow the embeddings to be adapted during the training phase. This way, the character embeddings are themselves able to capture relations between characters that commonly appear together.

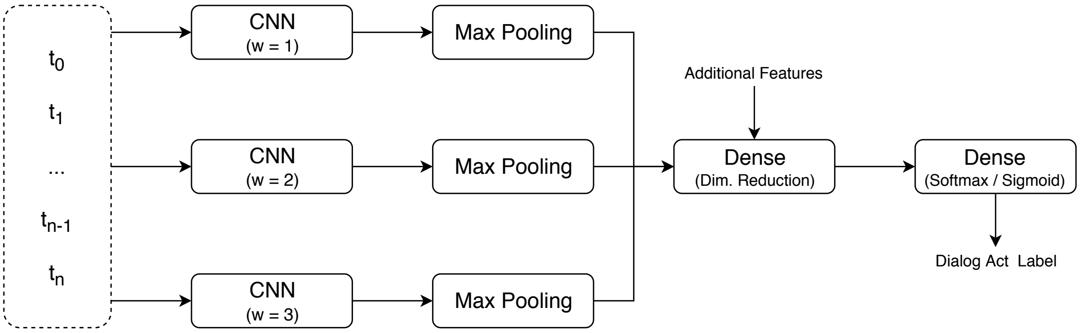

Figure 1. The generic architecture of the network used in our experiments. t_i corresponds to the embedding representation of the ith token. w corresponds to the context window size of the CNN. The number of parallel CNNs and the respective window sizes vary between experiments. Those shown in the figure correspond to the ones used by Liu et al. [6] in their experiments, which we adopt for the word-level baselines.

To generate the representation of the segment, the token embeddings are passed through a set of parallel temporal CNNs with different context window sizes followed by a max pooling operation. Each of these CNNs focuses on groups of characters of a certain size and, thus, it is important to use a set of context windows able to capture the different morphological aspects that are relevant for the task. In our experiments we explored windows of up to ten characters independently before selecting the best combination. The representation of the segment is then given by the concatenation of the output of each max pooling operation. This way, it contains information concerning all the captured patterns. Finally, to summarize and capture the most important information provided by the multiple parallel CNNs, the segment representation is passed through a dimensionality reduction layer. Additionally, this layer applies dropout during the training phase to reduce the probability of overfitting to the training data.

As shown in our previous study [17] and in that by Liu et al. [6], context information is highly important for the task, especially that provided by the preceding segments. Such information can be provided to the network by concatenating it to the representation of the segment before it is passed through the dimensionality reduction layer. However, since our study focused on the difference between using character- and word-level tokenization, we only included context information in a final experiment to assess whether it affects the character- and word-level approaches in the same manner and to compare our results on the SwDA corpus with those achieved by Liu et al. [6]. In that case, we provided the same context information as they did to achieve the best results reported in their study, that is, a flag stating whether the speaker changed and the concatenation of the one-hot encoding of the dialog act classifications of the three preceding segments. Concerning the latter, we used the gold standard annotations. Thus, the results we report for experiments that include context information provide an upper bound for the approach and the performance is expected to be slightly lower in a real scenario. However, as stated in Section 2, this approach for providing context information from the preceding segments surpasses other approaches even when using automatic classifications [6,17].

Finally, the reduced segment representation is passed through a dense layer to obtain its classification. In its original version, this dense layer performs the softmax activation function to identify the class with highest probability. We used the same approach in most of our experiments. The exception is when predicting the labels of the domain-dependent levels of the dialog act annotations of the DIHANA corpus, since they pose a multilabel classification problem. In that case, the dense layer performs the sigmoid activation function to obtain an estimation of the probability of each class. Then, those with values above 0.5 are activated.

To assess whether the character- and word-level approaches capture complementary information, we also performed experiments that combined both approaches. In that scenario, we used the architecture shown in Figure 2. In this case, two segment representations are generated in parallel, one based on the characters in the segment and other on its words. Those representations are then concatenated to form the final representation of the segment. The following steps do not differ from the architecture with a single branch. That is, context information can be added to the segment representation before it is passed to the dimensionality reduction layer and, subsequently, to the output layer.

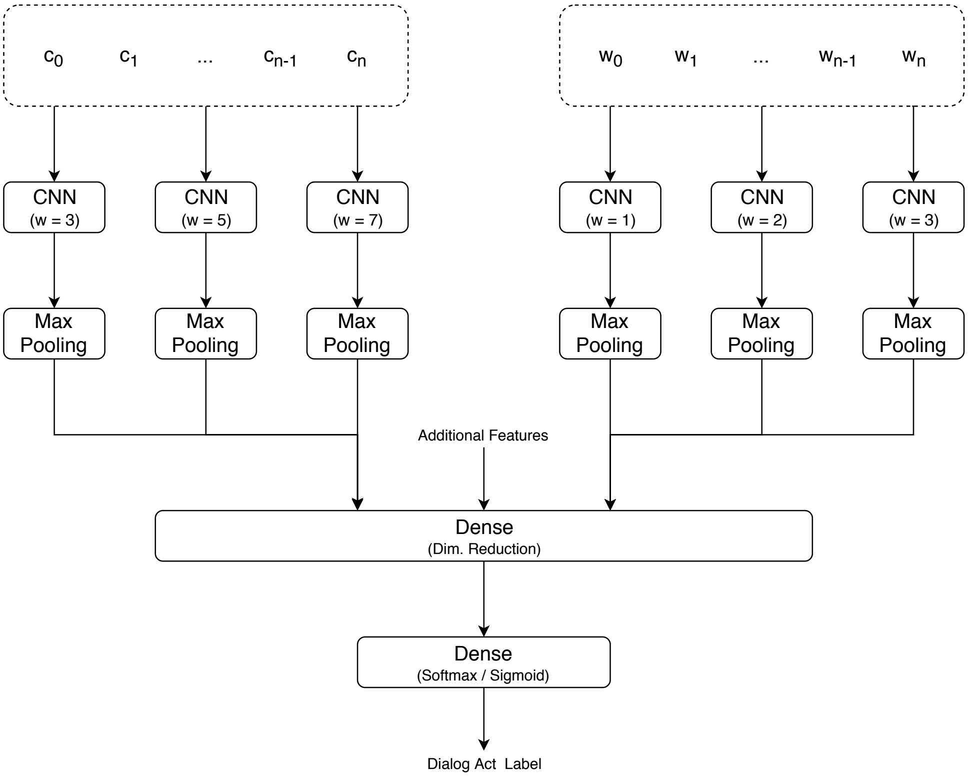

Figure 2. The architecture of the network that combines the character- and word-level approaches. c_i corresponds to the embedding representation of the ith character while w_i corresponds to the embedding representation of the ith word. The context window sizes of the CNNs in the character-level branch refer to those that achieved best performance in our experiments. Those in the word-level branch correspond to the ones used by Liu et al. [6] in their experiments.

Implementation Details

We implemented the networks used in our experiments using the Keras [36] interface to TensorFlow [37]. For performance reasons, we used the Adam optimizer [38] when updating the network weights. For the multilabel classification problems posed by the domain-dependent levels of the dialog act annotations of the DIHANA corpus, we used the binary cross entropy loss function. In every other case, we used the categorical cross entropy loss function. Each CNN applied 100 filters. The reduced segment representation generated by the dimensionality reduction layer was 200-dimensional.

To attenuate the influence of random initialization and the non-determinism of some operations run on Graphics Processing Unit (GPU), we performed ten runs of each experiment. In each run, the mini-batch size was 512 and the training phase stopped after ten epochs without improvement on the validation set.

4.3. Evaluation

To evaluate the performance of our approach, we required at least one metric, a partition of the dataset to evaluate on, and a baseline for comparison. All of these are presented below.

4.3.1. Evaluation Metric

In all the studies summarized in Section 2, the performance on dialog act recognition was evaluated in terms of accuracy. Thus, accuracy is also the metric we used in this study. It is important to note that accuracy is a highly penalizing metric for performance on the multilabel classification problems posed by the domain-dependent levels of the dialog act annotations of the DIHANA corpus, since it does not account for partial correctness [39]. However, due to space constraints and for a matter of consistency, we do not report results for specialized metrics. Since we performed ten runs of each experiment, the results presented in the next section refer to the average and standard deviation accuracy values in percentage form obtained over those runs.

4.3.2. Dataset Partition

There is a standard data partition of the SwDA corpus into a training set of 1115 conversations, a test set of 19 conversations, and a future use set of 21 conversations [15]. In our experiments, we used the latter as a validation set. However, it is important to note that not all previous studies on the SwDA used this partition.

The DIHANA corpus is partitioned into five dialog-based folds to be used for cross-validation [40]. Furthermore, it is important to note that, when predicting the labels of the domain-dependent levels, we did not consider the segments whose Level 1 label refers to discourse structuring functions or reveals problems in the dialog. We opted for this approach because those segments never have labels on the remaining levels and, thus, their correct classification is only dependent on the performance of the Level 1 classifier.

There is no standard partition of the LEGO corpus. Thus, we also split it into five folds at the dialog level and evaluated it using cross-validation.

The dialogs recorded during each phase of the VERBMOBIL project are organized into two different training, validation, and test sets. Additionally, the attribution to a set is not dialog-based, but rather speaker-based. Furthermore, there are some annotated segments that do not belong to any of the sets. Given these constraints, we opted for also splitting the German dialogs used in our experiments into five dialog-based folds and performing cross-validation.

4.3.3. Baselines

To assess the performance of the character-level approach in comparison to that of word-level approaches, we defined two baselines. Both use the CNN-based approach shown in Figure 1 with context window sizes one, two, and three and without appending context information to the representation of the segment. However, while one uses randomly initialized word embeddings that are adapted during the training phase, the other uses fixed pre-trained word embeddings. The pre-trained embeddings for English, Spanish, and German were obtained by applying Word2Vec [24] on the English Wikipedia (https://dumps.wikimedia.org/enwiki/) [23], the Spanish Billion Word Corpus [41], and the German Wikipedia (https://dumps.wikimedia.org/dewiki/) [42], respectively.

Additionally, we defined a third baseline that replicates the approach by Liu et al. [6]. It consists of the baseline with pre-trained embeddings combined with context information from three preceding segments in the form of their gold standard annotations and speaker change information in the form of a flag. In another study [43], we showed that, when predicting the dialog acts in the lower levels of the DIHANA corpus, context information from the upper levels is also important. Additionally, previous experiments on the LEGO corpus show that in that case only the system segment that immediately precedes each user segment is relevant to accurately predict its classification [31]. However, for a matter of consistency, we used the same context information for every dataset.

5. Results and Discussion

In this section, we present and discuss the results achieved in each of our experiments. In Tables 6–9, the columns labeled SwDA-V and SwDA-T refer to the results achieved on the validation and test sets of the SwDA corpus, respectively. The columns labeled DIHANA1, DIHANA2, and DIHANA3 refer to the results achieved when predicting the Level 1, 2, and 3 labels on the DIHANA corpus, respectively. The column labeled VM refers to the results achieved on the VERBMOBIL corpus and the column labeled LEGO is self-explanatory. The meaning of each row is explained in the caption of each table.

5.1. Baselines

Starting with the word-level baselines, in Table 6, we can see that using pre-trained embeddings led to improvements on English data. However, the same was not true for data in other languages. However, that is not due to language, but rather to the fact that both DIHANA and VERBMOBIL have fixed domains, while SwDA does not. Thus, while the randomly initialized embeddings could adapt for the first two, they could not do so for the latter and the performance on the validation and test sets was impaired if the embeddings Were overfit to the training data. Although the LEGO corpus also has a fixed domain, since the transcriptions of user segments were obtained automatically and have multiple problems, fixed pre-trained embeddings were useful in this case.

Table 6. Accuracy (%) results of the word-level baselines. The first row refers to the baseline using randomly initialized word embeddings. The second row refers to the baseline using pre-trained word embeddings (PT). The last row refers to the baseline including context information (Ctx).

	SwDA-V	**SwDA-T**	**LEGO**	**DIHANA1**	**DIHANA2**	**DIHANA3**	**VM**
Random	76.17 ± 0.19	72.23 ± 0.20	55.93 ± 0.10	91.96 ± 0.13	70.53 ± 0.24	97.23 ± 0.05	75.38 ± 0.07
PT	76.81 ± 0.32	73.11 ± 0.26	56.60 ± 0.10	91.98 ± 0.12	70.71 ± 0.33	96.01 ± 0.08	75.42 ± 0.08
PT + Ctx	81.29 ± 0.30	78.35 ± 0.36	86.71 ± 0.11	98.26 ± 0.04	91.80 ± 0.06	97.47 ± 0.04	76.43 ± 0.08

Using context information improved the results in every case, but at different scales. On the SwDA corpus, the improvement was in line with that reported by Liu et al. [6]. Our results differ from those reported in their paper mainly because they did not use the standard validation and test partitions of the dataset. The improvement on the Level 1 of the DIHANA corpus was higher than on the SwDA corpus simply because there are fewer labels. However, the highest improvements were achieved on the LEGO corpus and the Level 2 of the DIHANA corpus. The first case is explained by the nature of the tag set, which contains multiple variations of the same label according to the context (e.g., *Confirm Departure*, *Confirm Destination*, *Confirm Time*, amd *Confirm Bus*). The second case is explained by the nature of the dialogs, which feature multiple question–answer pairs. The segments in each pair focus on the same kind of information and, thus, have the same Level 2 labels. The improvement on the VERBMOBIL corpus was lower since we collapsed the labels into the top level of the tree, removing most of the cases that were ambiguous. On the Level 3 of the DIHANA corpus, there was only a slight improvement since that level focuses on the information that is explicitly referred to in the segment. Thus, in that case, context information from the preceding segments at the same level was typically irrelevant.

5.2. Character Windows

Regarding the character-level experiments, Table 7 shows the results achieved when considering sets of characters of different sizes on segments stripped of capitalization and punctuation information. As expected, considering each character individually led to the worst results, since the information concerning relations between characters was lost. By considering pairs of characters, the performance improved by at least 3% and up to 17%. Widening the window up to four or five characters kept improving the performance, but at a lower scale. Considering wider windows beyond that was typically harmful. However, similar to what Liu et al. [6] showed at the word level, different context windows could capture complementary information. Thus, it was beneficial to combine multiple windows. In our experiments, the best results were achieved using three context windows, which considered groups of three, five, and seven characters, respectively. The sizes of these windows were relevant, since the shortest window could capture most affixes in English and the small affixes in Spanish and German, the middle window could capture larger affixes and most lemmas, and the widest window could capture larger words, some agglutinative aspects of German, and inter-word information.

Table 7. Accuracy (%) results obtained using different character windows. The last row refers to the combination of windows that led to the best results in our experiments. The remaining rows refer to the performance of individual windows.

	SwDA-V	SwDA-T	LEGO	DIHANA1	DIHANA2	DIHANA3	VM
1	65.42 ± 0.17	60.81 ± 0.23	52.04 ± 0.34	85.71 ± 0.29	64.24 ± 0.25	76.16 ± 0.36	62.89 ± 0.19
2	72.21 ± 0.47	67.52 ± 0.54	55.28 ± 0.30	91.54 ± 0.14	70.09 ± 0.18	93.35 ± 0.24	74.34 ± 0.15
3	74.32 ± 0.55	70.00 ± 0.35	56.13 ± 0.24	92.17 ± 0.10	70.82 ± 0.35	95.29 ± 0.19	75.60 ± 0.08
4	74.56 ± 0.19	70.64 ± 0.49	56.17 ± 0.22	92.22 ± 0.14	71.16 ± 0.22	95.88 ± 0.07	76.31 ± 0.12
5	75.09 ± 0.52	70.91 ± 0.38	56.19 ± 0.19	92.28 ± 0.11	71.08 ± 0.29	96.05 ± 0.11	76.19 ± 0.10
7	75.35 ± 0.23	70.86 ± 0.34	55.92 ± 0.19	92.24 ± 0.13	71.18 ± 0.13	96.21 ± 0.11	76.03 ± 0.09
10	75.10 ± 0.36	70.97 ± 0.35	56.11 ± 0.15	92.16 ± 0.13	71.13 ± 0.16	96.12 ± 0.07	75.65 ± 0.10
(3, 5, 7)	76.08 ± 0.33	72.08 ± 0.42	56.92 ± 0.14	92.44 ± 0.12	71.42 ± 0.26	96.31 ± 0.17	77.22 ± 0.07

5.3. Segment Preprocessing

As mentioned in Section 3, we hypothesized that capitalization is not relevant for dialog act recognition. In Table 8, we can can see that the hypothesis holds for every dataset except DIHANA, as the results obtained when using capitalized segments did not significantly differ from those obtained using stripped segments. However, on the DIHANA corpus, and more specifically on Level 1, using capitalized segments led to an improvement of nearly 2%. Since this was not expected, we looked for the source of the improvement. By inspecting the transcriptions, we noticed that, contrarily to user segments, the system segments do not contain mid-segment capitalization. Thus, proper nouns, such as city names, which are common in the dialogs, are capitalized differently. Since only 5 of the 11 Level 1 labels are common to user and system segments, identifying its source reduced the set of possible dialog acts for the segment. Thus, the improvement observed when using capitalization information is justified by the cues it provides to identify whether it is a user or system segment.

Table 8. Accuracy (%) results obtained using different segment preprocessing approaches. The first row refers to segments stripped of punctuation and capitalization information. The second row refers to segments containing capitalization information, while the third row refers to segments containing punctuation. The fourth row combines both punctuation and capitalization information. The last row refers to lemmatized segments stripped of punctuation and capitalization information.

	SwDA-V	SwDA-T	LEGO	DIHANA1	DIHANA2	DIHANA3	VM
Stripped	76.08 ± 0.33	72.08 ± 0.42	56.92 ± 0.14	92.44 ± 0.12	71.42 ± 0.26	96.31 ± 0.17	77.22 ± 0.07
Caps (C)	76.04 ± 0.28	71.94 ± 0.26	56.92 ± 0.14	94.25 ± 0.15	71.41 ± 0.24	96.42 ± 0.07	77.03 ± 0.06
Punct (P)	76.85 ± 0.21	73.17 ± 0.32	56.92 ± 0.14	93.71 ± 0.07	71.50 ± 0.18	96.46 ± 0.10	79.76 ± 0.08
P + C	76.73 ± 0.25	73.14 ± 0.40	56.92 ± 0.14	95.48 ± 0.04	71.52 ± 0.22	96.47 ± 0.07	79.73 ± 0.07
Lemma	75.21 ± 0.27	71.40 ± 0.12	56.63 ± 0.11	92.39 ± 0.06	71.41 ± 0.23	96.40 ± 0.13	77.09 ± 0.08

In Table 8, we can also see that, as expected, punctuation provided relevant information for the task, especially when predicting domain-independent labels, leading to an improvement around one percentage point on the SwDA and DIHANA corpora and above 2.5% on the VERBMOBIL corpus. Punctuation was not as important to predict the labels of the domain-dependent levels of the DIHANA corpus since they are more keyword oriented. On the LEGO corpus, we cannot draw any conclusions concerning the influence of punctuation and capitalization information since the automatic transcription of user segments provided by the ASR system does not provide that information.

With the exception of the Level 3 of the DIHANA corpus, by including punctuation information, the character-level approaches could achieve results that are in line with or surpass the word-level baselines without context information. Furthermore, it is important to note that, while the improvement was negligible on the SwDA corpus, it was nearly 2% on the Level 1 of the DIHANA corpus and above

4% on the VERBMOBIL corpus. This suggests that the performance of the character-level approach in comparison with the word-level improved with the level of inflection of the language.

Since Level 3 is focused on the presence of certain words in the segment, it makes sense that the character-level approach did not perform as well, since considering the characters introduced unnecessary entropy. This was partially confirmed by the fact that, by using lemmatized segments, the results on Level 3 improved in comparison to the segments that consider multiple forms of the same lexeme. On the remaining cases, as expected, we observed a decrease in performance when using lemmatized segments. This proved that affixes were relevant for transmitting intention. However, in most cases, the decrease was slight, which suggests that most information concerning intention could be transmitted using a simplified language that does not consider variations of the same lexeme and that those variations were only relevant for transmitting some specific intentions.

5.4. Combinations

Finally, Table 9 shows the results obtained by combining the word- and character-level approaches, as well as when providing context information. We can see that, in every dataset except DIHANA, the combination of the word- and character-level approaches led to the best results, meaning that both approaches could capture complementary information. Furthermore, on the DIHANA corpus, the only exception was when predicting Level 2 labels. In this case, the best results were achieved by the character-level approach on its own. Overall, this suggests that information at the sub-word level was relevant for the task, independently of the language and the domain of the dialog.

Table 9. Accuracy (%) results obtained using combined information from multiple sources. The first two rows refer to the best results at the word and character levels, respectively. The third row refers to the combination of both approaches. The last three rows include context information (Ctx).

	SwDA-V	**SwDA-T**	**LEGO**	**DIHANA1**	**DIHANA2**	**DIHANA3**	**VM**
Word (W)	76.81 ± 0.32	73.11 ± 0.26	56.60 ± 0.10	91.98 ± 0.12	70.71 ± 0.33	97.23 ± 0.05	75.42 ± 0.08
Char (C)	76.85 ± 0.21	73.17 ± 0.32	56.92 ± 0.14	95.48 ± 0.04	71.52 ± 0.22	96.47 ± 0.07	79.76 ± 0.08
C + W	78.00 ± 0.16	74.01 ± 0.35	57.86 ± 0.16	95.68 ± 0.03	71.07 ± 0.20	97.35 ± 0.05	80.03 ± 0.09
W + Ctx	81.29 ± 0.30	78.35 ± 0.36	86.71 ± 0.11	98.26 ± 0.04	91.80 ± 0.06	97.47 ± 0.04	76.43 ± 0.08
C + Ctx	81.82 ± 0.26	78.41 ± 0.30	86.98 ± 0.13	98.75 ± 0.15	92.96 ± 0.06	96.59 ± 0.09	80.25 ± 0.09
All	82.00 ± 0.27	79.01 ± 0.16	87.24 ± 0.14	99.10 ± 0.04	91.90 ± 0.06	97.48 ± 0.06	80.67 ± 0.06

When including context information, the combination of the word- and character-level approaches still led to the best results, with the exception of the Level 2 of the DIHANA corpus. This means that, on the SwDA corpus, our approach surpassed the one by Liu et al. [6], which was the state-of-the-art approach at the time of our study. On the LEGO corpus, without considering context information, the 57.86% achieved by our approach surpassed the 52.40% achieved using SVMs [31]. When considering context information, the study using SVMs achieved 87.95% accuracy. However, context information was provided in the form of the words of the preceding system segment. On the DIHANA corpus, we improved the 97.92% achieved on Level 1 in our study at the word level [43]. On the lower levels, the comparison is not as straightforward since we used additional context information in that study. However, considering our results, we expect that the 94.38% achieved on Level 2 in that study could be improved by introducing character-level information. On Level 3, the 97.48% reported in this article surpassed the 96.34% achieved in that study. However, the improvement was not due to the use of character-level information, but rather the use of adaptable word embeddings. Our results on the VERBMOBIL corpus could not be compared with those of previous studies, such as that by Reithinger and Klesen [44], since they used a different label set.

6. Conclusions

In this article, we have assessed the importance of information at a sub-word level, which cannot be captured by word-level approaches, for automatic dialog act recognition in three different languages—English, Spanish, and German—on dialogs covering multiple domains, and using both domain-independent and domain-dependent tag sets.

We used character-level tokenization together with multiple character windows with different sizes to capture relevant morphological elements, such as affixes and lemmas, as well as long words and inter-word information. Furthermore, we have shown that, as expected, punctuation is important for the task since it is able to provide both direct and indirect cues regarding intention. On the other hand, capitalization is irrelevant under normal conditions.

Our character-level approach achieved results that are in line or surpass those achieved using state-of-the-art word-based approaches. The only exception was when predicting the domain-specific labels of Level 3 of the DIHANA corpus. However, this level refers to information that is explicitly referred to in the segment. Thus, it is highly keyword oriented and using character information introduces unnecessary entropy. In the remaining cases, the character-level approach was always able to capture relevant information, independently of the domain of the dialog, the domain-dependence of the dialog act labels, and the language. Concerning the latter, it was interesting to observe that the highest performance gain in comparison with the word-level approaches occurred in German data, while the lowest occurred in English data. This suggests that the amount of relevant information at the sub-word level increases with the level of inflection of the language.

Furthermore, our experiments revealed that in most cases the character- and word-level approaches capture complementary information and, consequently, their combination leads to improved performance on the task. In this sense, by combining both approaches with context information, we achieved state-of-the-art results on SwDA corpus, which is the most explored corpus for dialog act recognition. Additionally, given appropriate context information, our approach also achieved results that surpass the previous state-of-the-art on the DIHANA and LEGO corpora. On the VERBMOBIL corpus, we were not able to compare our results with those of previous studies, since they use different label sets.

In terms of morphological typology, although English has a more analytic structure than Spanish, and German has some agglutinative aspects, the three are fusional languages. Thus, as future work, it would be interesting to assess the performance of the character-level approach when dealing with data in analytic languages, such as Chinese, and agglutinative languages, such as Turkish. However, it is hard to obtain annotated corpora with such characteristics and it is hard to draw conclusions without knowledge of those languages.

Author Contributions: Data Curation: E.R.; Formal Analysis: E.R., R.R., and D.M.d.M.; Investigation: E.R.; Methodology: E.R.; Software: E.R.; Supervision: R.R. and D.M.d.M.; Validation: E.R.; Writing—Original Draft: E.R.; and Writing—Review and Editing: E.R., R.R., and D.M.d.M.

Funding: This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2019 and by Universidade de Lisboa.

Acknowledgments: We would like to thank José Miguel Benedí for providing us access to the DIHANA corpus.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ASR	Automatic Speech Recognition
CMU	Carnegie Mellon University
CNN	Convolutional Neural Network
DNN	Deep Neural Network

DRLM	Discourse Relation Language Model
DSTC4	Dialog State Tracking Challenge 4
FCT	Fundação para a Ciência e a Tecnologia
GloVe	Global Vectors for Word Representation
GPU	Graphics Processing Unit
HMM	Hidden Markov Model
IVR	Interactive Voice Response
LSTM	Long Short-Term Memory
MRDA	ICSI Meeting Recorder Dialog Act Corpus
NLP	Natural Language Processing
NN	Neural Network
POS	Part-of-Speech
RNN	Recurrent Neural Network
RNNLM	Recurrent Neural Network Language Model
SVM	Support Vector Machine
SwDA	Switchboard Dialog Act Corpus
WER	Word Error Rate
WoZ	Wizard of Oz

References

1. Searle, J.R. *Speech Acts: An Essay in the Philosophy of Language*; Cambridge University Press: Cambridge, UK, 1969.
2. Král, P.; Cerisara, C. Dialogue Act Recognition Approaches. *Comput. Inform.* **2010**, *29*, 227–250.
3. Kalchbrenner, N.; Blunsom, P. Recurrent Convolutional Neural Networks for Discourse Compositionality. In Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality, Sofia, Bulgaria, 9 August 2013; pp. 119–126.
4. Lee, J.Y.; Dernoncourt, F. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2016), Diego, CA, USA, 12–17 June 2016; pp. 515–520.
5. Ji, Y.; Haffari, G.; Eisenstein, J. A Latent Variable Recurrent Neural Network for Discourse Relation Language Models. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2016), Diego, CA, USA, 12–17 June 2016; pp. 332–342.
6. Liu, Y.; Han, K.; Tan, Z.; Lei, Y. Using Context Information for Dialog Act Classification in DNN Framework. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 2160–2168.
7. Santos, C.D.; Zadrozny, B. Learning Character-level Representations for Part-of-Speech Tagging. In Proceedings of the 31st International Conference on Machine Learning (ICML 2014), Beijing, China, 21–26 June 2014; pp. 1818–1826.
8. Jaech, A.; Mulcaire, G.; Hathi, S.; Ostendorf, M.; Smith, N.A. Hierarchical Character-Word Models for Language Identification. In Proceedings of the International Workshop on Natural Language Processing for Social Media, New York, NY, USA, 11 July 2016; pp. 84–93.
9. Zhang, X.; Zhao, J.; LeCun, Y. Character-level Convolutional Networks for Text Classification. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; Volume 1, pp. 649–657.
10. Ribeiro, E.; Ribeiro, R.; de Matos, D.M. A Study on Dialog Act Recognition using Character-Level Tokenization. In Proceedings of the 18th International Conference on Artificial Intelligence: Methodology, Systems, Applications, San Francisco, CA, USA, 2–7 July 2018; pp. 93–103.
11. Jurafsky, D.; Shriberg, E.; Biasca, D. *Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual*; Technical Report Draft 13; Institute of Cognitive Science, University of Colorado: Colorado Springs, CO, USA, 1997.

12. Benedí, J.M.; Lleida, E.; Varona, A.; Castro, M.J.; Galiano, I.; Justo, R.; de Letona, I.L.; Miguel, A. Design and Acquisition of a Telephone Spontaneous Speech Dialogue Corpus in Spanish: DIHANA. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, 22–28 May 2006; pp. 1636–1639.
13. Schmitt, A.; Ultes, S.; Minker, W. A Parameterized and Annotated Spoken Dialog Corpus of the CMU Let's Go Bus Information System. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, 21–27 May 2012.
14. Kay, M.; Norvig, P.; Gawron, M. *VERBMOBIL: A Translation System for Face-to-Face Dialog*; University of Chicago Press: Chicago, IL, USA, 1992.
15. Stolcke, A.; Coccaro, N.; Bates, R.; Taylor, P.; Van Ess-Dykema, C.; Ries, K.; Shriberg, E.; Jurafsky, D.; Martin, R.; Meteer, M. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Comput. Linguist.* **2000**, *26*, 339–373. [CrossRef]
16. Gambäck, B.; Olsson, F.; Täckström, O. Active Learning for Dialogue Act Classification. In Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH), Florence, Italy, 27–31 August 2011; pp. 1329–1332.
17. Ribeiro, E.; Ribeiro, R.; de Matos, D.M. The Influence of Context on Dialogue Act Recognition. *arXiv* **2015**, arXiv:1506.00839.
18. Manning, C.D. Computational Linguistics and Deep Learning. *Comput. Linguist.* **2015**, *41*, 701–707. [CrossRef]
19. Goldberg, Y. A Primer on Neural Network Models for Natural Language Processing. *J. Artif. Intell. Res.* **2016**, *57*, 345–420. [CrossRef]
20. Kim, S.; D'Haro, L.F.; Banchs, R.E.; Williams, J.; Henderson, M. The Fourth Dialog State Tracking Challenge. In *Dialogues with Social Robots*; Springer: Singapore, 2016; pp. 435–449.
21. Janin, A.; Baron, D.; Edwards, J.; Ellis, D.; Gelbart, D.; Morgan, N.; Peskin, B.; Pfau, T.; Shriberg, E.; Stolcke, A.; et al. The ICSI Meeting Corpus. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hong Kong, China, 6–10 April 2003; pp. 364–367.
22. Mikolov, T.; Karafiát, M.; Burget, L.; Cernocký, J.; Khudanpur, S. Recurrent Neural Network Based Language Model. In Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH), Chiba, Japan, 26–30 September 2010; pp. 1045–1048.
23. Khanpour, H.; Guntakandla, N.; Nielsen, R. Dialogue Act Classification in Domain-Independent Conversations Using a Deep Recurrent Neural Network. In Proceedings of the 26th International Conference on Computational Linguistics (COLING), Osaka, Japan, 11–16 December 2016; pp. 2012–2021.
24. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, 5–10 December 2013; pp. 3111–3119.
25. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods on Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
26. Godfrey, J.J.; Holliman, E.C.; McDaniel, J. SWITCHBOARD: Telephone Speech Corpus for Research and Development. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), San Francisco, CA, USA, 23–26 March 1992; Volume 1, pp. 517–520.
27. Carletta, J. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Comput. Linguist.* **1996**, *22*, 249–254.
28. Rotaru, M. *Dialog Act Tagging Using Memory-Based Learning*; Technical Report; University of Pittsburgh: Pittsburgh, PA, USA, 2002.
29. Webb, N.; Ferguson, M. Automatic Extraction of Cue Phrases for Cross-corpus Dialogue Act Classification. In Proceedings of the 23th International Conference on Computational Linguistics (COLING), Beijing, China, 23–27 August 2010; pp. 1310–1317.
30. Raux, A.; Bohus, D.; Langner, B.; Black, A.W.; Eskenazi, M. Doing Research on a Deployed Spoken Dialogue System: One Year of Let's Go! Experience. In Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH), Pittsburgh, PA, USA, 17–21 September 2006; pp. 65–68.

31. Chorianopoulou, A.; Palogiannidi, E.; Iosif, E.; Koutsakis, P.; Georgiladakis, S.; Trancoso, I.; Batista, F.; Moniz, H.; Ribeiro, E.; Abad, A.; et al. *SpeDial D2.1 Interim Report on IVR Analytics and Evaluation*; Technical Report 2.1; SpeDial Consortium; 2015. Available online: https://sites.google.com/site/spedialproject/risks-1 (accessed on 3 March 2019).
32. Alcácer, N.; Benedí, J.M.; Blat, F.; Granell, R.; Martínez, C.D.; Torres, F. Acquisition and Labelling of a Spontaneous Speech Dialogue Corpus. In Proceedings of the 10th International Conference Speech and Computer (SPECOM), Patras, Greece, 17–19 October 2005; pp. 583–586.
33. Martínez-Hinarejos, C.D.; Sanchis, E.; García-Granada, F.; Aibar, P. A Labelling Proposal to Annotate Dialogues. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC), Las Palmas, Spain, 29–31 May 2002; Volume 5, pp. 1566–1582.
34. Jekat, S.; Klein, A.; Maier, E.; Maleck, I.; Mast, M.; Quantz, J.J. *Dialogue Acts in VERBMOBIL*; Technical Report; DFKI: Kaiserslautern, Germany, 1995.
35. Alexandersson, J.; Buschbeck-Wolf, B.; Fujinami, T.; Kipp, M.; Koch, S.; Maier, E.; Reithinger, N.; Schmitz, B.; Siegel, M. *Dialogue Acts in VERBMOBIL-2*, 2nd ed.; Technical Report; DFKI: Kaiserslautern, Germany, 1998.
36. Keras: The Python Deep Learning Library. 2015. Available online: https://keras.io/ (accessed on 2 March 2019).
37. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: https://www.tensorflow.org/ (accessed on 2 March 2019).
38. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
39. Sorower, M.S. *A Literature Survey on Algorithms for Multi-Label Learning*; Technical Report; Oregon State University: Corvallis, OR, USA, 2010.
40. Martínez-Hinarejos, C.D.; Benedí, J.M.; Granell, R. Statistical Framework for a Spanish Spoken Dialogue Corpus. *Speech Commun.* **2008**, *50*, 992–1008. [CrossRef]
41. Spanish Billion Words Corpus and Embeddings. 2016. Available online: http://crscardellino.me/SBWCE/ (accessed on 2 March 2019).
42. GermanWordEmbeddings. 2015. Available online: https://github.com/devmount/GermanWordEmbeddings (accessed on 2 March 2019).
43. Ribeiro, E.; Ribeiro, R.; de Matos, D.M. Hierarchical Multi-Label Dialog Act Recognition on Spanish Data. *Linguamática* **2019**, *11*, under review.
44. Reithinger, N.; Klesen, M. Dialogue Act Classification Using Language Models. In Proceedings of the Fifth European Conference on Speech Communication and Technology, Rhodes, Greece, 22–25 September 1997; pp. 2235–2238.