Two-level classification for dialogue act recognition in task-oriented dialogues

Philippe Blache[1], **Massina Abderrahmane**[2], **Stéphane Rauzy**[1], **Magalie Ochs**[3], and **Houda Oufaida**[2],
[1] LPL-CNRS-AMU, Aix-en-Provence, France (firstname.last@univ-amu.fr)
[2] ESI, Algiers, Algeria (h_oufaida@esi.dz)
[3] LIS-CNRS-AMU, Marseille, France (firstname.last@univ-amu.fr)

Abstract

Dialogue Act classification becomes a complex task when dealing with fine-grain labels. Many applications require such level of labelling, typically automatic dialogue systems. We present in this paper a 2-level classification technique, distinguishing between *generic* and *specific* dialogue acts (DA). This approach makes it possible to benefit from the very good accuracy of generic DA classification at the first level and proposes an efficient approach for specific DA, based on high-level linguistic features. Our results show the interest of involving such features into the classifiers, outperforming all other feature sets, in particular those classically used in DA classification.

1 Introduction

Task-oriented dialogue systems are specific in many respects. First, and this is the most important characteristic, they focus on a limited semantic domain, rendering the comprehension operation closer to *slot filling* than deep semantic understanding. Moreover, these applications correspond to very specific interactions, usually with one of the speakers (the human or the machine) leading the dialogue. In such situations, recognizing dialogue acts (DA) becomes extremely useful as a preliminary step of the understanding process: associating speaker's utterances to DAs makes it possible to identify very efficiently the type of information they bear and the knowledge to be conveyed. Many works on dialogue act classification have been done for a long time (Stolcke et al., 2000). The first question consists in defining the set of relevant dialogue acts for the system. Different generic DA annotation schemes have been proposed, including an ISO standard (Bunt et al., 2012; Bunt et al., 2017). In addition, several dialogue corpora have been annotated and made available, among which two are particularly used: *Switchboard Dialog Act Corpus, (SwDA)* (Jurafsky et al., 1997) and *Meeting Recorder Dialog Act (MRDA)* (Shriberg et al., 2004). These datasets have allowed to train different classifiers. In these works, two preliminary questions arise: the identification of the dialogue acts to be classified and the features on which the classifier must be based. In most cases, the classification task targets a limited set of classes (corresponding to very general DAs such as *statement*, *question* etc.) and features remain at a low level (n-grams of words, characters, word embeddings, etc.). The performance of these classifiers is generally very good. However, the proposed task (i.e. the targeted tagset) remains often too general for an optimal use in a dialog system. Moreover, and this is a recurring problem in this type of approach, it is sometimes difficult to interpret the results and understand precisely the relative impact of the different features on the model.

In this paper, we address the question of DA classification for helping comprehension and dialogue supervision in the context of task-oriented dialogue systems. Several works have shown the importance of DAs for guiding dialogues, in particular in the context of adaptive and socially-aware systems used for training social skills (Zhao et al., 2016). In such cases, the dialogue system plays the role of a person to whom information must be transmitted (Ochs et al., 2018b). The targeted information, the semantic context, as well as the way the information should be transmitted are known in advance by the system. We are then in the situation where the most difficult task for the system is to generate an appropriate reaction to the way the human has delivered the information: appropriate feedbacks, clarification questions, surprise, emotional reactions, etc. This is the reason why DA classification plays an important role, more than a classical understanding or intent/slot-filling (Firdaus et al., 2019).

The identification of dialogue acts can therefore be an extremely efficient pre-processing for comprehension, provided that DAs are specific enough to guide the system efficiently. DA identification comes to a classification task which requires three preliminary steps: identifying an appropriate tagset, collecting an adequate corpus and annotating it. The annotation stage is an important issue, this task being done mainly manually. As a consequence, the corresponding datasets remain rather small, which can be problematic for applying machine learning

Proceedings of the 28th International Conference on Computational Linguistics, pages 4915–4925
Barcelona, Spain (Online), December 8-13, 2020

techniques. Moreover, as described in this paper, an efficient DA classification makes use of a large number of features. One way to address this question of high dimensionality is to study precisely the features to be involved in the model, their characteristics and behavior.

Another characteristics of dialogue systems is that comprehension has to be done incrementally and in real time. When modeling a dialogue based on natural dialogue corpora, the global set of speakers' utterances are taken into account, making it possible to analyze large sequences. However, for automatic dialogue systems, the only utterances containing information to be processed are those already produced by the human speaker. This has important consequences on the technique to be applied: it cannot take into account a dialogue in its entirety and then cannot be bi-directional (no possible forward looking when calculating the system's answer): the model should mainly/only rely on the user's utterances (utterances produced by the system being limited to speaker's reactions). We show in this study that DA classification for these specific dialogue situation can be based only on user's utterances, offering the possibility of an online and real time processing.

This paper proposes a general methodology to address the problem of classifying specific dialogue acts (in place of generic DAs) based on a set of interpretable features, adapted to the characteristics of data available for this type of task. We propose in a first part a summary of the existing techniques and current results in DA classification. In the second section, we describe the specific data we are working on and the constitution of the dataset. The rest of the article is devoted to the presentation and discussion of the results.

2 Dialog act classification

Dialogue acts (DAs) correspond to communicative functions describing the illocutionary force (speaker's intention) that can be associated with utterances in the discourse. DA classification consists in associating to each segment of the discourse (that can be a turn, an inter-pausal unit or a segment returned by a speech recognition system) a label corresponding to the communicative function. DA classification is a classical problem for which many propositions have been done (Stolcke et al., 2000; Ang et al., 2005; Tavafi et al., 2013; Lee and Dernoncourt, 2016; Chen et al., 2018; Kumar et al., 2017; Raheja and Tetreault, 2019). Two preliminary questions must be addressed in this task before a classification technique can be applied: the choice of the tagset, and the identification of the features used by the classifiers. We propose in this section to quickly address these issues before presenting the main results available in the literature.

The DA tagset: The study of dialogue acts has led to several annotation schemes. Among these, the *Dialogue Act Markup in Several Layers (DAMSL)* (Core and Allen, 1997), which served as a basis for the annotation of two reference corpora, *SwDA* and *MRDA*, cited above. More recently, the *DIT++* schema, which led to the establishment of the *ISO 24617-2* standard (Bunt et al., 2012; Bunt et al., 2017), proposed an organization based on different dimensions (e.g. turn management, social obligations management, etc.) each containing different dialogue acts. In total, *DIT++* proposes to distinguish more than 100 dialogue acts. *DAMSL*, on the other hand, proposes 226 dialogue acts, which are generally clustered into 42 labels. From an automatic classification perspective, too many classes do not lead to efficient results. Therefore, many studies propose to limit the number of classes by using either generic metaclasses or the most frequently used. For example, several works are based on the *DAMSL* tagset reduced to 5 classes (statements, questions, backchannels, fillers and disruptions) (Ang et al., 2005). Some other works, such as the reference one (Stolcke et al., 2000) or the recent state of the art described in (Chen et al., 2018), use a tagset reduced to the 5 most frequent tags (*statement, backchannel, opinion, abandoned, agreement*).

Reducing the tagset can also be done for adjusting the classification to the needs of a specific domain dialog system. This is for example the case in (Anakina and Kruijff-Korbayova, 2019), describing communication in the context of robot-assisted disaster response. In this work, a specific set of ISO metaclasses adapted to the needs of the system is proposed, clustering the 20 most useful dialogue acts into 8 metaclasses (`Contact, Inform, Affirmative, Request, Question, Confirm, Disconfirm, Negative`).

The features: The second question to be answered for DA classification is that of the features used by the classifiers. Most works use low-level features, such as n-grams of characters and words, length of the utterance, etc. (Stolcke et al., 2000; Kim et al., 2016; Sennrich et al., 2017). These features are usually complemented by distributional information such as word embeddings (Chen et al., 2018; Kumar et al., 2017; Tran et al., 2017; Chakravarty et al., 2019), which can be weighted with TF-IDF (Joulin et al., 2017; Raheja and Tetreault, 2019). In several studies, this information is also complemented by linguistic features such as prosody or morpho-syntactic information (Shriberg et al., 1998; Stolcke et al., 2000; Bothe et al., 2018; Tran et al., 2017).

Related works: Many different approaches have been tested for DA classification. The following table summarizes the main results recently obtained in this task, indicating the type of classification technique, the dataset and the accuracy. Two of these studies represent the state of the art for the *MRDA* an the *SwDA* datasets. In a recent paper, (Chen et al., 2018) has proposed a new approach based on *CRF-Attentive Structured Network*. This technique relies on a hierarchical representation distinguishing three levels (words, utterances, conversation), integrating a

Authors	Algorithm	Corpus (and tagset)	Accuracy
(Grau et al., 2004)	Naïve Bayes	Switchboard corpus	66%
(Stolcke et al., 2000)	HMM with a trigram language model	Switchboard (5 most frequent DAs)	71%
(Shen and Lee, 2016)	Attentional RNN	Switchboard	72.6%
(Kalchbrenner and Blunsom, 2013)	Recurrent Convolutional Neural Network	Switchboard	73.9%
(Tavafi et al., 2013)	SVM-HMM	Switchboard (16 DAs)	74.32%
(Bothe et al., 2018)	Context learning (3 preceding DAs) + of RNNs	Switchboard	77.34%
(Chen et al., 2018)	CRF-Attentive Structured Network	Switchboard (5 DAs)	81.3%
(Raheja and Tetreault, 2019)	CRF decoding, contextual attention	Switchboard	82.9%
(Tavafi et al., 2013)	SVM-HMM	MRDA dataset (11 DAs)	80.5%
(Lendvai and Geertzen, 2007)	Naïve Bayes	MRDA	82%
(Chen et al., 2018)	CRF-Attentive Structured Network	MRDA (5 DAs)	91.7%

Table 1: State of the art in DA classification

memory to take into account contextual dependencies. At the first layer, a fine-grained representation (integrating in particular character and word-level embeddings, POS, named entity). Each utterance is encoded with a bi-directional GRU (*Gated Recurrent Unit*), implementing the context influence (capturing long term dependencies across the conversation). An attention mechanism (selecting the most relevant information) is integrated, introducing weights based on the similarity between the input memory (obtained from the word embedding layer) and the current utterance. This technique has been evaluated for the classification of the 5 most frequent DAs in *SwDA* and *MRDA*, obtaining respectively an accuracy of 81.3% and 91.7% (still the state-of-the-art for *MRDA*).

Finally, (Raheja and Tetreault, 2019) describes the state-of-the art DA classification evaluated against *SwDA*. They propose a combination of techniques already presented above (CRF decoding, contextual attention, and character-level word embeddings) complemented with self-attentive representation learning. They report an accuracy of 82.9% in the identification of the 43 classes of *SwDA*, where the mean accuracy of other comparable methods is 75% on the same dataset.

3 The dataset

The goal of this paper is to design a DA classification method providing inputs for the understanding module of a dialogue system. We give first details about the dataset and the use case before describing the annotation process.

3.1 Use case: training doctors

We propose to focus on a specific application: *using a dialogue system for training doctors to break bad news* (Ochs et al., 2018b). This work is part of the ACORFORMED project (`http://www2.lpl-aix.fr/~acorformed/`) which consists in asking trainees, following a given scenario, to announce the patient a problem that occurred during a medical act. In such a context, the dialogue structure is very specific. First, the semantic domain is closed and the system has the entire knowledge of the scenario. Moreover, the way the doctor announces the new has to follow some recommendations (Schnebelen et al., 2011), in particular by structuring the discourse in different phases in the announcement (greetings, presentation of the context, description of the problem and its solutions, etc.). In such situations, the interlocutor roles are not balanced: the doctor (in our case the human trainee) is the main speaker and the patient (played by the dialogue system) reacts to the doctor's utterances, without taking the lead of the conversation.

We collected a corpus of training sessions, in French, organized between doctors (the trainee) and patients (played by human actors). The corpus is made of 7 sessions each lasting around 15mns. The audio input (representing 37,000 words) has been transcribed and manually corrected. The corpus has been automatically segmented into *inter-pausal units* (with pauses higher than 250ms). Each inter-pausal unit forms an utterance, 1,822 such utterances have been produced throughout the 7 dialogues by the doctors.

3.2 Annotation scheme

Several proposal have been done in terms of dialog acts annotation. In particular, based on *DIT++* , ISO 24617-2 scheme (Bunt et al., 2012; Bunt et al., 2017) proposes a hierarchy of such acts. In our work, DA classification is conceived as a pre-processing step before understanding. Our goal is therefore to try identifying as precise DA classes as possible. As explained above, the dialogue in this type of discourse is precisely structured. Besides the classical opening and closing phases, the main part of the dialogue consists in different stages in the information transfer: the patient's initial state (having justified the hospitalization), the bad new description (typically an incident during a surgery) and the patient's current state. Moreover, the doctor also gives explications, asks questions,

Anonymous_name	*DAMSL*	ISO
Opening	*Conventional-opening*	Opening, Initial greeting
Init_state	*Statement-non-opinion*	Inform
Init_remediation	*Statement-non-opinion*	Inform
Bad_new_state	*Statement-non-opinion*	Inform
Bad_new_remediation	*Statement-non-opinion*	Inform
Current_state	*Statement-non-opinion*	Inform
Current_remediation	*Statement-non-opinion*	Inform
Reassurance	*Statement-non-opinion*	Social obligations management functions, Empathy expression
Explication	*Summarize/reformulate*	Inform
Social_interaction	*Apology, Offers, Options, Commits*	Social obligations management functions, Empathy expression, Apologizing, Expressing gratitude
Discourse	*Uninterpretable, Affirmative non-yes answers, Reject, Other*	Feedback Functions; Dialogue structuring functions
Question	*Info-Request*	Question
Closing	*Conventional-closing*	Dialogue closing

Table 2: DA tagset correspondence with DAMSL and ISO.

reassures the patient and have different types of social interactions.The following table lists the complete set of dialogue acts used in the *ACORFORMED* project and their correspondence with *DAMSL* and *ISO* labels. Note that the selected dialog acts are based on a linguistics analysis of the dialogs of the corpus described in (Ochs et al., 2018a). In this sense, the tagset is specific to the needs of the project but remains generic to the class of task-oriented systems in the context of dialogues aiming at transferring information from one speaker to the other. The *ACORFORMED* dialogue acts are semantically fine-grained, which is of great help for the understanding process, but represents a much harder problem for automatic classification: we have in this case more classes, with few differences between them (compare for example the different classes describing the state of the patient). Note that we do not integrates in this scheme a label Other, taking into account the specificity of the application domain.

3.3 Corpus annotation

The corpus has been manually annotated by 5 annotators among which two experts. The Fleiss kappa inter-rater coefficient has been applied and shows a inter-annotator agreement of 0.518, which corresponds to a fair agreement according to usual standards (Fleiss et al., 2003). The following table shows the agreement by class, which gives an idea of the different levels of difficulty:

	Kappa		Kappa		Kappa
Init_remediation	0.479	Current_remediation	0.500	Opening	0.739
Init_state	0.349	Current_state	0.434	Reassurance	0.388
Bad_new_state	0.440	Discourse	0.750	Social_interaction	0.641
Bad_new_remediation	0.313	Explication	0.274	Closing	0.688

The high number of classes mechanically decreases the Fleiss kappa coefficient. However, we consider the result to be acceptable. It also underlines the difficulty of the annotation task (as well as that of classification) at this level of precision. Figure 1 illustrates the class distribution, showing its unbalanced nature: in the corpus, the class current_state represents 21% of the total number, whereas the class closing corresponds to 0.5%. This is obviously one of the problems we will have to fix when using machine learning techniques.

4 Feature selection

We propose to explore different approaches for classifying DA in the context of our specific dataset. More precisely, we will work specifically on two aspects: the feature set (by adding high-level features on top of the classical ones) and classification techniques (by comparing a classical multiclass approach with a specific case of hierarchical classification). As it is usually the case, a preliminary step of data cleaning is applied, suppressing special characters, accents, stopwords and normalizing the font case. A separate pre-processing has been done, adding lemmatisation to these operations. Moreover, we used *MarsaTag* (Rauzy et al., 2014), a lexical processing tool integrating different functionalities (including POS tagging) in order to extract the lexical features.

Classical features : We first experiment classification with a set of features classically used for DA identification. It consists in combining *TF-IDF* principles with word and character n-grams. Applying a principal component analysis, we extracted 4 combinations to be tested:

Morpho-syntactic features : We propose to explore the role of low-level morpho-syntactic features, based on POS categories:

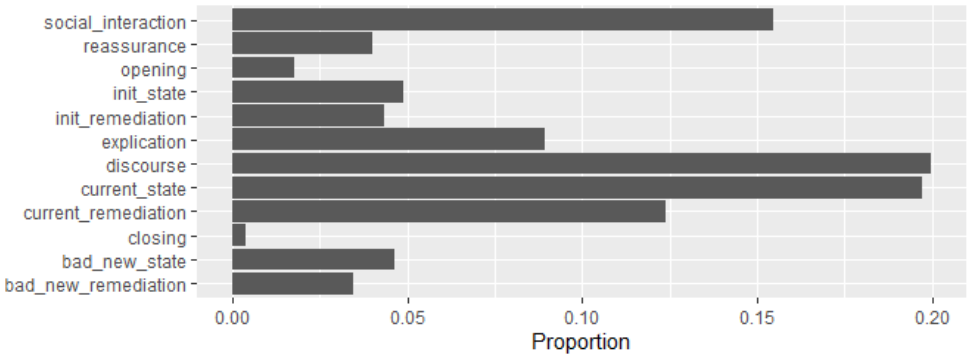

Figure 1: *The* 4 *dominant classes* `discourse`, `current_state`, `social_interaction`, `current_remediation` *cover* 78% *of the corpus,* 20% *is made of the* 5 *intermediate classes* `reassurance`, `init_state`, `init_remediation`, `bad_new_remediation` *and* `bad_new`. *The remaining* 2% *of the corpus is composed of the* 2 *low occurrence classes* `opening` *and* `closing`.

f-TFIDF	Classical TFIDF on word n-grams (from 1 to 3 words, keeping the 250 best) and character n-grams (from 3 to 5 chars, keeping the 250 best), then a total of 500 features.
s-TFIDF	The f-TFIDF features, filtered with a singular value decomposition in order to obtain a better representation density
w-TFIDF	TFIDF only based on the word n-grams, keeping the 500 best.
l-TFIDF	TFIDF based on the lemmas n-grams, keeping the 500 best

Table 3: Classical features

Lexical features : We created a dictionary specific to our domain, containing medical words in which we distinguished pathological terms vs. others. Moreover, we annotated the data with a specific label tagging the medical words depending on they appear for the first time in the dialogue or not (corresponding to the given/new distinction used in discourse analysis). From this information, the following features are extracted:

Context features : As proposed in several works (Bothe et al., 2018; Raheja and Tetreault, 2019), context (i.e. the labels of the preceding dialogue acts) is taken into account. We implemented three different context representations, in a 1 to 5 window: *one hot encoding* of the preceding DAs, *bag-of-words* (encoding the number of times the DA appears in the context of the utterance), *n-grams* of words (up to 0.5% frequency). The following features are used: `Prec_onehot1`, `Prec_onehot2`, `Prec_bow2`, `Prec_bow5`, `Prec_ngram3` and `Prec_ngram_1+3` (feature `Prec` concatenated with `Prec_ngram3`).

Syntactic features : High-level syntactic information can play a role in the characterization of certain classes. In particular, dialogue sequences corresponding to a description or an explanation are usually associated to more complex structure, with more modifiers (adjectives and adverbs) and more complex clauses (subordinates, relatives, prepositional phrases). We propose two features for a simple approximation of these characteristics:

5 Flat classification

Our choice of the classification techniques takes into account two aspects: the size of the dataset and our intention to have interpretable models. These two goals can be contradictory. Having a rather small dataset increases the risk of overfitting, which leads to chose the simplest models. On the other hand, in order to improve interpretability, we want to integrate as many linguistic features as possible. We chose in this perspective to test different simple machine learning models: *Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Decision tree, Random Forest*; some of them are known to be particularly adapted to learn on small dataset (Forman and Cohen, 2004). Different regularization techniques has been applied according to the model: *L1, L2, Elasticnet*, as well as tuning the margin parameter for SVM or the tree depth for random forests. Feature selection is then applied using an Anova to select the k-best features. The best classifier is retrained with this new feature set. Two classical methods for hyperparameter tuning, *RandomSearch* and *GridSearch*, have been applied on the training set, with k-cross validation. Considering the size of the dataset, a k-fold cross-validation has been applied, with $k = 5$, in order to have a validation set large enough to be statistically significant (different values of k have been tested). Moreover, as shown in the previous section, class distribution is imbalanced, some of them representing 0.3% of the data, where others correspond to 20%. We used *Synthetic Minority OverSampling Technique* (SMOTE) which consists in adding new points by combining the points of the minority class with the closest neighbors. We refined this method by filtering the output with *Tomek Link Removal*, removing the points that are the closest neighbors of each other. As a result, different tests have been done on imbalanced and balanced classes (with the same class

DM	Number of discourse markers in the utterance
FP	Number of filled-pauses
TOK	Number of tokens

Table 4: Morpho-syntactic features

MED	Nb of medical terms
MED_P	Nb of pathological terms
MED_O	Nb of non-pathological terms
MED_New	Nb of *new* medical terms
MED_Given	Nb of *given* medical terms (i.e. that already occurred in the dialogue)

MED-P_New	Nb of *new* pathological terms
MED-P_Given	Nb of *given* pathological terms
MED-O_New	Nb of *new* non-pathological terms
MED-O_Given	Nb of *given* non-pathological terms *given*

Table 5: Lexical features

proportion for the training and the test sets).

Evaluation metrics are accuracy and balanced accuracy, which gives an idea of the mean performance by class. We chose to use accuracy (and not other metrics such as weighted accuracy) in order to compare our results with the literature. Finally, the corpus contains 1,822 utterances, split in 80% for the training set, 20% for the test set.

In this first experiment, we classified directly, with a unique classifier, the 13 classes: `init_state`, `init_remediation`, `bad_new`, `bad_new_remediation`, `current_remediation`, `current_state`, `opening`, `closing`, `discourse`, `explication`, `question`, `reassurance`, `social_interaction`. Several tests have been done with different feature combinations (note that Clit_i stands for clitic 1st to 3rd person):

Table 8 shows the raw results, from which we can compare the different proposals we have. The best results (73.8% accuracy, 63% balanced accuracy) are obtained with XGBoost applied to the complete set of features. Note that Random Forest led to comparable accuracy. As it is the case with other works on DA classification, using context (in our case a simple one-hot encoding of one preceding DA) always significantly improves the results for all classifiers (improvement of 15 points in the case of SVM). What is new in this study is that involving the all set of linguistic features also significantly improves accuracy: comparing the best classifier trained with the classical features and the context (column C+CONT) with that trained with all features shows an improvement of 2 points, reaching 72% (keeping a good balanced accuracy of 59%). Linguistic features approximate high-level syntactic information about discourse elaboration and the use of complex structures. DA classes are very different from each others with this respect: explaining a difficult situation requires higher linguistic complexity than asking a question or having a social interaction. This results is in line with our goal to go towards an interpretable model.

6 Two-level classification

Multiclass classification is always prone to several difficulties. Moreover, the identification of classes corresponding to fine-grained knowledge (e.g. state, remediation, etc.) used in the perspective of dialogue understanding is much more difficult than other types of classification. Addressing this issue can be done thanks to *hierarchical classification* (Silla and Freitas, 2011). Chaining classifiers in particular when there is a lot of classes to consider can be extremely helpful by breaking down the problem into small simpler problems, increasing the overall performance. As explained above, at the difference with DA classification triggering meta-classes, we need in our task a fine-grained classification, which is by definition (and also due to the specificity of our dialogue classes) harder to obtain. In order to fit with our final application (utterance pre-processing for spoken language understanding in a dialogue), we propose to take advantage both from the robustness of meta-class classification and the interest of the fine-grained one. We are then in a specific top-down classification, limited to two levels (DA meta-classes and leaf classes). This corresponds to a "*Local Classifier per Level*" approach which consists of training a multi-class classifier for each level (Freitas and de Carvalho, 2008). This method is known to have two main drawbacks. First, as it is the case with other top-down class-prediction approaches, errors at the higher level are propagated downwards. Second, this method ignores parent-child class relationships. However, it presents in our case an important advantage: information provided at each level will be used directly for utterance understanding in the dialogue system.

We propose to keep as first-level classes the metaclasses specified in the ISO 24617-2 scheme (see section 4): `Opening`, `Discourse`, `Inform`, `Question`, `Closing`. The second level details the Inform metaclass in 8 subclasses: `Init_state`, `Init_remediation`, `Bad_new_state`, `Bad_new_remediation`, `Current_state`, `Current_remediation`, `Social_interaction`, `Explication`.

6.1 Level-1 classification

We tested different feature configurations. This section focuses on an overview of the best results we obtained. We compare for that three different feature combinations: *CLASS* (*f-TFIDF*, to which an principal component analysis

MODIFIERS	The ratio of the number of adjectives and adverbs to the total number of tokens in the utterance: $\frac{nb\_adj+nb\_adv}{\sum tokens}$
CLAUSES	The ratio of the number of conjunctions, pronouns and prepositions to the total number of tokens: $\frac{nb\_conj+nb\_prep+nb\_pro}{\sum tokens}$

Table 6: Syntactic features

CLASS	Classical features (*f-TFIDF*)
C+MED	*CLASS* **plus** lexical features + *MED* + *MED_New*
C+MED+LING	*C+MED* **plus** *DM, FP, TOK, MODIFIERS, CLAUSES, nb_N, nb_V, nb_aux, nb_Clit*$_i$
C+CONT	CLASS **plus** Prec_onehot1
C+MED+LING+CONT	*C+MED+LING* **plus** Prec_onehot1
C+LING+CONT+MED_PO	*C+LING+Prec_onehot1* **plus** *MED_P, MED_O, MED-P_New, MED-O_New*

Table 7: Feature sets

has been applied in order to keep the 350 best features among the initial 500), *C+MED* (*CLASS* plus medical lexical features MED and MED_New) and *C+MED+LING* (the previous linguistic features). The following table reports the results obtained for the different classifiers. The linear regression classifier with an Anova to select the k-best features leads to the best results. As expected, the accuracy is very high, reaching 94% (89% of balanced accuracy). Note that the 1st-level classes are relatively stable and easy to recognize, the context feature did not bring any improvement there. It is out of the scope of this paper to compare these results with the state of the art in similar DA annotations: datasets and language are totally different. However, remind that the higher performance on the *SwDA* corpus is 81.3% and 91.7% on the *MRDA*.

Taking into account the size of the dataset, we have tried oversampling methods, as described in the previous section. Table 10 reports the results obtained for the best classifiers, using the *C+MED+LING* feature set, showing that oversampling does not improve the results obtained by the linear regression model.

6.2 Level-2 classification

The second step of the classification consists in applying a new classifier to the sequences labeled Inform in the first step. The Inform sub-clases are: `Init_state, init_remediation, bad_new, bad_new_remediation, current_state, current_remediation, explication, social_interaction, reassurance`.

We have tested different feature configurations and focus her on the following feature selection: *CLASS (f-TFIDF), C+MED+LING, C+CONT, C+MED+LING+CONT*. We also tried a new feature combination, noted *C+MED+LING+CONT+MEDPO*, adding to *C+MED+LING+CONT* the specific medical features distinguishing pathological terms from others (*MED-P_New, MED-P_Given, MED-O_New, MED-O_Given*).

Note that, as for the flat classification technique, applying imbalanced classes pre-processing does not improves the results, reason why they are not reported here.

7 Reducing the number of classes

As explained above, we have tried in our approach to use fine-grained classes, at a precise semantic level. This is of course extremely interesting in the perspective of helping the understanding mechanisms of a dialogue system. But of course, this introduce a new level of difficulty. We have tried to reduce this complexity by defining more general classes, factorizing some of the dialogue acts used previously. In this experiment, we used the following classes:

- State: this new class regroups the init_state, bad_new_state and current_state.
- Remediation: gathering init_remediation, bad_new_remediation and current_remediation
- Social_interaction: gathering social_interaction and reassurance
- Opening: gathering greeting, presentation and object
- Closing, Discourse, Question: same definition as in the previous typology

Several clustering experiments have been done, without leading to interesting results (typically clusters with the same word, not really useful in our task) or even more problematically difficult to interpret (in the case of clustering based on linguistic features). We applied the same techniques and feature sets as previously described and obtained the following results:

Remind that this experiment concerns a classifier for 7 classes. It is therefore not possible to compare directly these results with those of the 13 classes classifiers. The decision of the choice for the final classifier depends then on the classes granularity level we consider optimal, taking into account the performance. The 13 classes flat classifier accuracy is 73.7% (63% balanced accuracy). On its site, the level-2 classification (with the 8 finer-grained classes) has an accuracy of 77.2% (71% balanced accuracy). This is to be compared, all things equal, with

	Eval.	CLASS	C+MED	C+M+LING	C+CONT	C+M+L+CONT	C+M+L+C+M_PO
LR	Acc.	0.5499	0.5227	0.5613	0.6838	0.6980	0.6980
	Bal.	0.3367	0.3362	0.3486	0.5466	0.5811	0.5587
SVM	Acc.	0.5299	0.5328	0.5328	0.7037	0.6838	0.7009
	Bal.	0.3333	0.3396	0.3295	0.5932	0.5289	0.5939
KNN	Acc.	0.4188	0.4330	0.4160	0.4160	0.4416	0.4387
	Bal.	0.2713	0.2820	0.2849	0.2908	0.3222	0.3126
Decision Tree	Acc.	0.3960	0.4046	0.5014	0.5726	0.6724	0.6752
	Bal.	0.2513	0.2053	0.2940	0.4436	0.5046	0.5043
Random Forest	Acc.	0.4758	0.5071	0.5442	0.6239	0.7293	0.7322
	Bal.	0.2288	0.3088	0.3221	0.4406	0.5569	0.5613
K-best feat.	Acc.	0.5499	0.5527	0.5613	0.7037	0.7293	0.7350
+ best classifier	Bal.	0.3367	0.3362	0.3486	0.5932	0.5533	0.5647
XGBoost	Acc.	/	/	/	/	0.7322	0.7379
	Bal.	/	/	/	/	0.6250	0.6308

Table 8: Flat classification results

	Eval.	CLASS	C+MED	C+MED+LING
LR	Acc.	0.8842	0.8875	0.9003
	Bal.	0.8319	0.8329	0.8668
SVM	Acc.	0.8875	0.8907	0.8842
	Bal.	0.8410	0.8610	0.8616
KNN	Acc.	0.7460	0.7781	0.7974
	Bal.	0.7632	0.7735	0.7844
Decision Tree	Acc.	0.8328	0.7974	0.8232
	Bal.	0.6779	0.6220	0.5791
Random Forest	Acc.	0.8553	0.8617	0.8714
	Bal.	0.6358	0.6166	0.6248
K-best feat.	Acc.	0.9035	0.9068	0.9421
+ best classifier	Bal.	0.8478	0.8638	0.8900

Table 9: Level-1 classification results

this 7-classes flat classifier which has an accuracy of 78.3% (59.6% balanced accuracy). We obtain then rather comparable results, the 7-classes having a slightly better accuracy (but lower balanced accuracy).

8 Discussion

Classifying DA is in itself a difficult task when trying to take into account a fine level of precision (therefore a high number of classes). Most of the experiments reported in the literature focus on general dialog acts, typically the 43 classes for *SwDA* that correspond to their general function played by the utterance in the dialogue (question, assessment, backchannel, etc.). When using such general DAs, automatic classification reaches easily a good accuracy. However, the task becomes difficult when entering into finer-grained DAs (required when using such classification as a pre-processing step for dialogue systems). In this case, classes correspond to a specific informational content much more difficult to identify than classical functional dialogue acts (in particular because of the natural variability of language). The complexity of this task is reflected in the medium level of inter-annotator agreement we obtained.

The first strategy we explored consists in classifying directly our 13 classes. Our results are rather good, reaching an accuracy of 73.8%, which is not so far from the average accuracy of the simpler functional classification (75%). *XGBoost* with a feature set involving, on top of n-grams, higher level linguistic, lexical and syntactic features led to the best performance in comparison with 6 other classifiers. These results confirm the importance of the context (in our case the previous DA), which improves by 12 points other feature sets. They also show the interest of taking into account linguistic features. Lexical features, which are domain-dependent (frequency of medical terms, distinction between pathological terms from others), improve by 2 points the results obtained with contextual features. But note that the *MED* feature set also contains discourse-level information by integrating the count of new referents in the utterance. This feature plays an important role by distinguishing the phases describing patient's state from others: new terms are mainly introduced during the description of the initial state of the patient as well as the description of the bad new. On their side, higher level linguistic features (richer lexical description with the *Modifiers* feature, more complex syntactic constructions with *Clauses* feature) also improves by 2 more points the results with the context and the terminological features. As explained above, these features play a role by distinguishing sequences during which language is more elaborated than others. Typically, descriptive or explicative utterances fall into this category. This is an interesting result not only because it improves the performance, but also because we can interpret directly the role of such features in the model.

	Eval.	Base	Oversampling	SMOTE
LR	Acc.	0.9003	0.8617	0.8489
	Bal.	0.8668	0.8595	0.8530
K-best feat.	Acc.	0.9421	0.9325	0.9325
+ best classifier	Bal.	0.8900	0.8869	0.8869

Table 10: Level-1 oversampling

	Eval.	CLASS	C+MED+LING	C+CONT	C+MED+LING+CONT	C+M+L+C+MEDPO
LR	Acc.	0.4054	0.4363	0.6911	0.7066	0.7143
	Bal.	0.2626	0.3151	0.6431	0.6317	0.6317
SVM	Acc.	0.4131	0.4517	0.7413	0.7336	0.7336
	Bal.	0.2827	0.3151	0.6431	0.6317	0.6317
KNN	Acc.	0.3629	0.3398	0.5869	0.5830	0.5946
	Bal.	0.2978	0.2760	0.4507	0.4544	0.4689
Decision Tree	Acc.	0.3205	0.3475	0.7104	0.7104	0.7104
	Bal.	0.1813	0.1902	0.6569	0.6373	0.6400
Random Forest	Acc.	0.3745	0.3900	0.7413	0.7606	0.7722
	Bal.	0.2439	0.2607	0.6496	0.6782	0.7101
K-best feat.	Acc.	0.4093	0.4979	0.7452	0.7606	0.7568
+ best classifier	Bal.	0.2635	0.2833	0.6579	0.6782	0.6951

Table 11: Level-2 classification results

The second strategy we explored in this paper, relying on a hierarchical classification in two steps, shows the interest of mixing generic and specific classes. The results obtained in the first step, classifying 5 generic DAs, outperform the state-of-the art with an accuracy at 94.2% (obtained by the linear regression model with k-best feature selection) and a balanced accuracy at 89%. In this task, the context feature cannot be used directly, taking into account the class factorization. We can see that linguistic features, as for flat classification, play a primordial role, outperforming in all cases the other feature sets. This is due to the fact that the majority class, `Inform`, gathers all classes possibly impacted by lexical richness and syntactic complexity. Let's note moreover that these very good results are very interesting in the perspective of dialogue control: 3 of the 5 classes (`Opening`, `Discourse`, `Closing`) can be directly used in order to generate a response of the dialogue system without entering into a specific processing.

The second step of the classification consists in refining the `Inform` class, identifying classes corresponding at the same time to dialogue acts and semantic information (states, remediation, social interaction and explication). In this case again, Random forest gives the best classifier. As with all other types of classification, involving linguistic features leads to the best results, which confirms the interest of such feature set whatever the type of dialogue act classification. The final accuracy reaches 77.2% (balanced accuracy 71%) improving the average level of fine-grained DA classification.

9 Conclusion

DA classification usually relies on low-level features. We have shown in this paper that bringing high level linguistic features in the model improves all types of classification (flat or hierarchical). In particular, we have shown that this approach outperforms the state-of-the-art for generic DA classification (corresponding to the level-1 classification in our experiment). Even more interestingly, we have shown that this feature set also leads to a very good accuracy for the more complex task of fine-grain DA classification. These results open the way to new solutions for language understanding in the context of task-oriented dialogue systems: classifying fine-grained DAs offers the possibility to precisely identify the semantic frames to be used in the semantic representation of the dialogue.

Acknowledgements

This work, carried out within the Institut Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government, managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

References

T. Anakina and I. Kruijff-Korbayova. 2019. Dialogue act classification in team communication for robot assisted disaster response. In *SIGDIAL Meeting on Discourse and Dialogue*, pages 299–410.

J. Ang, Y. Liu, and E. Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. *ICASSP*, pages 1061–1064.

	Eval.	CLASS	C+MED+LING	C+CONT	C+MED+LING+CONT	C+M+L+C+MEDPO
LR	Acc.	0.6122	0.6616	0.7034	0.7072	0.7148
	Bal.	0.3882	0.4185	0.4301	0.4524	0.4566
SVM	Acc.	0.6160	0.6426	0.6768	0.7072	0.7072
	Bal.	0.3937	0.4568	0.4598	0.5177	0.5177
KNN	Acc.	0.5323	0.5513	0.5361	0.5551	0.5551
	Bal.	0.3379	0.4890	0.3404	0.3526	0.3526
Decision Tree	Acc.	0.5247	0.5894	0.6616	0.7414	0.7414
	Bal.	0.3644	0.3621	0.5177	0.6105	0.5895
Random Forest	Acc.	0.5970	0.6388	0.7186	0.7643	0.7529
	Bal.	0.3441	0.5321	0.4398	0.5163	0.5102
K-best feat.	Acc.	0.6236	0.6426	0.7110	0.7719	0.7605
+ best classifier	Bal.	0.3911	0.4048	0.4365	0.5209	0.5145
XGBoost	Acc.	/	/	/	0.7833	0.7795
	Bal.	/	/	/	0.5960	0.5963

Table 12: Classification results with factorized classes

C. Bothe, C. Weber, S. Magg, and S. Wermter. 2018. A context-based approach for dialogue act recognition using simple recurrent neural networks. In *LREC 2018*.

H. Bunt, Jan A., Jae-Woong Choe, A. Chengyu Fang, K. Hasida, V. Petukhova, A. Popescu-Belis, and D. R. Traum. 2012. Iso 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 430–437.

H. Bunt, A. Chengyu Fang, and V. Petukhova. 2017. Revisiting the iso standard for dialogue act annotation. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annota- tion (ISA-13)*.

S. Chakravarty, R. Chava, and E. Fox. 2019. Dialog acts classification for question-answer corpora. In *Workshop ASAIL 2019*.

Z. Chen, R. Yang, Z. Zhao, D. Cai, and X. He. 2018. Dialogue act recognition via crf-attentive structured network. In *SIGIR*, pages 225–234.

M. Core and J. Allen. 1997. Coding dialogs with the damsl annotation scheme. In *Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35.

M. Firdaus, A. Kumar, A. Ekbal, and P. Bhattacharyya. 2019. A multi-task hierarchical approach for intent detection and slot filling. *Knowledge-Based Systems*, 183.

Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik, 2003. *Statistical Methods for Rates and Proportions (Third)*, chapter The Measurement of Interrater Agreement, pages 598–626. John Wiley & Sons, Inc.

G. Forman and I. Cohen. 2004. Learning from little: Comparison of classifiers given little training. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 161–172. Springer.

Alex Freitas and Andre de Carvalho. 2008. A tutorial on hierarchical classification with applications in bioinformatics. *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*, pages 119–145.

S. Grau, E. Sanchis, M. J. Castro, and D. Vilar. 2004. Dialogue act classification using a bayesian approach. In *9th Conference Speech and Computer SPECOM 2004*.

A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of EACL*.

D. Jurafsky, E. Shriberg, and D. Biasca. 1997. Switchboard swbd-damsl shallow-discourse- function annotation coders manual. Technical report, University of Colorado at Boulder Technical Report 97-02.

N. Kalchbrenner and P. Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. In *Workshop on Continuous Vector Space Models and their Compositionality*, pages 119–126.

Y. Kim, Y. Jernite, D. Sontag, and A. Rush. 2016. Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*, pages 2741–2749.

H. Kumar, A. Agarwal, R. Dasgupta, S. Joshi, and A. Kumar. 2017. Dialogue act sequence labeling using hierarchical encoder with crf. In *CoRR*.

J. Y. Lee and F. Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In *NAACL-HLT*.

P. Lendvai and J. Geertzen. 2007. Token-based chunking of turn-internal dialogue act sequences. In *SIGDIAL Workshop on Discourse and Dialogue*.

M. Ochs, P. Blache, G. Montcheuil, J-.M. Pergandi, R. Bertrand, J. Saubesty, D. Francon, and D. Mestre. 2018a. The acorformed corpus: Investigating multimodality in human-human and human-virtual patient interactions. In *CLARIN Annual Conference 2018*, page 16.

M. Ochs, D. Mestre, G. de Montcheuil, J.-M. Pergandi, J. Saubesty, E. Lombardo, D. Francon, and P. Blache. 2018b. Training doctors' social skills to break bad news: Evaluation of the impact of virtual environment displays on the sense of presence. *Journal on Multimodal User Interfaces*.

V. Raheja and J. Tetreault. 2019. Dialogue act classification with context-aware self-attention. In *NAACL 2019*.

S. Rauzy, Montcheuil G., and P. Blache. 2014. Marsatag, a tagger for french written texts and speech transcriptions. In *Second Asia Pacific Corpus Linguistics Conference*.

C. Schnebelen, F. Pothier, and M. Furney. 2011. Annonce d'un dommage associé aux soins. Technical report, Haute Autorité de Santé.

R. Sennrich, B. Haddow, and A. Birch. 2017. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics*.

S. Shen and H. Lee. 2016. Neural attention models for sequence classification: Analysis and application to key term extraction and dialogue act detection. In *INTERSPEECH 2016*, pages 2716–2720.

E. Shriberg, R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech. *Language and Speech*, 41(3-4):439–487.

E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. Technical report, International Computer Science Inst. Berkeley.

C. Silla and A. Freitas. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22:31–72.

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

M. Tavafi, Y. Mehdad, S. R. Joty, G. Carenini, and R. T. Ng. 2013. Dialogue act recognition in synchronous and asynchronous conversations. In *SIGDIAL Conference*, pages 117–121.

Q. Tran, I. Zukerman, and G. Haffari. 2017. A hierarchical neural model for learning sequences of dialogue acts. In *Conference of the EACL*.

R. Zhao, T. Sinha, A. W. Black, and J. Cassell. 2016. Automatic recognition of conversational strategies in the service of a socially-aware dialog system. In *SIGDIAL Conference*, pages 381–392.