The Role of Affect Analysis in Dialogue Act Identification

Nicole Novielli and Carlo Strapparava

Abstract—We present a qualitative analysis of the lexicon of Dialogue Acts: we explore the relationship between the communicative goal of an utterance and its affective lexicon as well as the salience of specific word classes for each speech act. Thought not constituting any deep understanding of the dialogue, automatic dialogue act labeling is a task that may be relevant for a wide range of applications in both human-computer and human-human interaction. The experiments described in this paper fit in the scope of a research study whose long-term goal is to build an unsupervised classifier that simply exploits the lexical semantics of utterances to automatically annotate dialogues with the proper speech acts.

Index Terms—Affective Lexicon, Dialogue Acts Recognition, Empirical Methods, Latent Semantic Analysis, Lexical Semantics.

1 Introduction

Affect is a fundamental issue of the communication process and has been widely studied in psychology and behavior science, as it constitutes a fundamental component of the human nature. According to Liu et al. [1], a successful social interaction is realized through a successful affective communication. Since the work of Austin [2] and Searle [3] on speech acts, the language has been seen as the primary indicator of people's attentional focus and communicative intention. Moreover, people use language also as an indicator of emotionality, even when no actual feelings are being reported or experienced by the speaker. Still, emotional words may be employed to convey other communicative intentions. It is the case of 'expressive' dialogue acts (i.e. apologizing, thanking or expressing sympathy) where affective language is often employed to simply represent and convey psychological attitudes [2], [4].

The study presented here fits in the scope of a research about automatic Dialogue Act (DA) recognition: our long-term goal is to define an unsupervised approach for automatically annotating natural dialogues with the proper speech acts by relying on empirical methods that simply exploit lexical semantics [5]. In particular, we present a study aimed at verifying whether it is possible to exploit affect analysis and affective language resources for improving the performance of our unsupervised approach. The recognition of actual emotions of the speaker during a dialogue (either in human-human or human-computer interaction) is out of the scope of this study. We rather intend to exploit state of the art techniques on affective language modeling and existing linguistic resources to improve our DA classifier.

The paper is organized as follows. In the next section, we provide motivation and background on speech act recognition and affect modeling in natural dialogue interfaces. In Section 3 we describe our unsupervised approach and illustrate how this study has been inspired by the findings of our previous experiments on DA recognition. In Section 4 we investigate the role of affective lexicon in utterances conveying specific dialogue acts; we use three different approaches involving three different linguistic resources to address our research questions and to improve the performance our unsupervised approach. We conclude discussing future work directions.

- N. Novielli is with the Department of Computer Science, University of Bari Aldo Moro, Bari, Italy. E-mail: novielli@di.uniba.it
- C. Strapparava is with the FBK-irst, Istituto per la Ricerca Scientifica e Tecnologica, Povo (Trento), Italy. E-mail: strappa@fbk.eu

2 Motivation and Background

The study presented in this paper has been inspired by the findings of our previous research on automatic labeling of natural language dialogues with the proper dialogue acts [6]. Rather than improving the performance of affect recognition from text, the goal of this study is to investigate to the role of affect analysis in DA identification. Our interest in the study of speech acts is motivated by the fact that they constitute the basis of everyday conversations. DA can be identified with the communicative goal of a given utterance [2] (e.g. asking for information, stating facts, expressing opinions, agreeing or disagreeing with the interlocutor) and there is a large number of applications that could benefit from automatic DA annotation, such as dialogue systems, blog analysis, automatic meeting summarization, user profiling by mean of dialogue pattern analysis, and so on.

In our previous research, we have defined an approach for DA recognition relying on empirical methods that exploit lexical semantics of sentences. It is based on the definition of set of *seeds* (words), which capture the specific semantic of each DA (i.e. its 'illocutionary force' [2]). The method has been evaluated in both a supervised and an unsupervised framework [5]. The findings of our experiments on DA recognition (see Section 3) support our intuition about the importance played by lexical semantics in dialogue act recognition while the analysis of misclassified cases, highlights how the main cause of error is the confounding between statements and opinions. By analyzing the utterances in our corpus, we observed that the more evident difference between objective statements and statement-like utterances expressing opinions is the wider use of slanted lexicon in the latter. Another problem is the poor recognition of 'behabitives' [2], i.e. typical situations in which no real feelings are expressed but still emotional words are employed to convey other communicative intentions (as for the use of 'I'm sorry' for apologizing or 'I'm glad you did it' for thanking). This particular class of speech acts has not been extensively studied in literature even if interesting logical formalization has been provided, using a BDI-like approach for the representation of the mental states they convey [4], that is an approach grounded on the Belief-Desire-Intention paradigm traditionally employed for modeling agents' behavior.

These observations are consistent with the findings of psychological studies demonstrating how a person's affective or psychological state considerably influence her language [7], [8]. Combining these observations with the theoretic and empirical literature about speech acts and the use of language in communication, we decide to perform a study aimed at investigating the relation between the affective lexicon of a given utterance and its communicative goal (i.e. its DA label). Hence, we formulate the following research questions:

(RQ1) *Does a relation exist between the affective lexicon and the communicative goal of an utterance?*

(RQ2) *Does the affective language analysis play a role in dialogue act identification?*

To address RQ1we apply lexical similarity techniques to textual input, consistently with the approach used in our previous research (see Section 4.1). In particular, we performed a qualitative study of the lexicon aimed at further investigating the relationship between the DA and the affective load of a given utterance, as well as the role played by lexical categories and their salience with respect to each DA. The results of this qualitative investigation were quite encouraging and have been exploited to address RQ2 (see Section 4.2).

A first attempt to exploit affective information in dialogue act disambiguation has been made by Bosma and André [9], with promising results. In their study, the recognition of emotions is based on sensory inputs that evaluate physiological user input. Users' expressions of emotion are also considered in the study described in [10]: facial expressions of users are exploited in an affect-enriched dialogue act classifier in tutorial dialogues. A discourse model has been implemented in the research described in [11], to exploit discourse relations for opinion polarity classification. In this study, an annotation scheme is used for labeling opinions, their polarities and their targets. The segmentation unit is the Dialogue Act and opinion annotations are mapped to the DA units in the corpus. Experimental results in both, a supervised and unsupervised framework, show how discourse-related features can be successfully exploited in solving misclassifications in opinion mining.

What we share with the studies described above is the underlying idea that a relation exist between discourse features and subjectivity in lexicon used by people in everyday conversations.

2.1 Related work on automatic DA recognition

In literature, automatic DA recognition has been treated as a task of text-classification, that is the problem of assigning a category label to a linguistic object. Samuel et al. [12] report one of the best performance, using a rule-based approach on textual features (accuracy of 71.22%). Rules exploit textual cues (i.e. words or text fragments) and are developed starting from the tagged VERMOBIL collection of 800 German and English conversations (18 DA labels), using a minimum entropy approach. The performance raises to 75.12% if the model is enriched with contextual features. On the same corpus, Reithinger and Klesen [13] combine a bayesian approach with *uni-* and *bi-* grams, achieving 67.18% and 74.7% of correctly classified labels for German and English, respectively. Stolcke et al. [14] achieve an accuracy of around 70% and 65% respectively on transcribed and recognized words by combining a discourse grammar, formalized in terms of Hidden Markov Models, with evidences about lexicon and prosody. They exploit the Switchboard corpus [15] of spontaneous task-free English conversations (see Section 3). The corpus is labelled according to the 42 tags of the SWBD-DAMSL schema and the recognition task is performed on six higher level classes of DA. The Switchboard corpus has been used also in studies addressing simultaneous dialog act segmentation and classification using Conditional Random Fields using lexical features, reporting an accuracy of 70% on the joint task[16]. A partially supervised framework has also been explored [17], using five broad classes of DA and obtaining an accuracy of about 79%.

Regardless of the model they use (discourse grammars, models based on word sequences or on the acoustic features or a combination of all these) the mentioned studies are developed in a supervised framework. Though, it is not always easy to collect large training material, partly because of manual labeling effort. In addition, with the advent of the Web, a large amount of written data about computer-mediated communication became available, raising the attractiveness of empirical methods of analysis in NLP. Hence, rather than improving the performance of supervised approaches, the long term goal of our research is to define DA lexical profiles that can be used in an unsupervised[1] framework for automatic

1. Or minimally supervised, since providing hand-specified seeds can be regarded as a minimal sort of supervision.

labeling of natural dialogues with the proper speech acts. We simply exploit text: even if prosody and intonation surely play a role (e.g. [14], [18]), nonetheless language and words are what the speaker uses to convey the communicative message and are just what we have at disposal when we consider texts found on the Web.

Unlike previous research, we aim at exploiting rules underlying human communication, rather than conventions. To Searle, speaking a language means to engage in a rule-governed form of behavior [3], as for playing a game. Searle describes the example of the chess game imaging that it is played in different countries: regardless of specific conventions (e.g. the way the king is represented), people can be said to play the same game all over the world, since they make game moves according to the same rules. Analogously, performing speech acts (the "moves" of the communication "game") in a given language means to behave according to rules underlying communication and the differences in surface realization arise because of different conventions. According to this vision, the fact that in French one can greet by saying *salut* and in English one can do it by saying *hi* is a matter of convention. But the fact that an utterance of a greeting device (i.e. the use of the words *salut* and *hi* in the two examples) under appropriate conditions (e.g. the participants to the dialogue just met or are about to start a conversation) counts as a greeting is a matter of rules, regardless of the specific language.

The fact that verbal communication is intentional and is a matter of rules and not (only) a matter of convention constitutes the main assumption underlying our research towards the definition of an unsupervised language-independent approach for DA categorization. We investigate the lexical semantics of dialogue moves to capture the rules underlying the surface realization of each DA. This vision is supported also by the analysis of the state of the art on speech act recognition, demonstrating how the structure of a discourse [19] and the communicative intentions of the speaker are reflected in linguistic realization of dialogue utterances [20]. Lexical cues, in particular, have been shown to play an important role as discourse markers: the lexical dimension is the one we investigate in our research. Moreover, as an original contribution, in this study we describe how affective language may play a role in enhancing the performance of an unsupervised framework for DA annotation.

2.2 Affect in natural language interaction

Researchers have widely recognized the importance of affect in communication and affective computing is now an established discipline [21]. In recent years, we have assisted to the flourishing of several projects aimed at classifying people's verbal and non behavior according to a set of either continuous or discrete emotional states or to the estimated value of some basic components of affective states (such as valence or intensity). This is due to the wide range of application domains in which affect has a fundamental role and for which affective state modeling could make a valuable and interesting contribution.

2.2.1 Affect Recognition in HCI

As far as Human-Computer Interaction is concerned, future intelligent interfaces are expected to embed some forms of emotional intelligence to enhance their naturalness and effectiveness. The integration of models of possibly many human cognitive capabilities, including affect analysis and generation, is becoming paramount and researchers have widely recognized the importance of studying the role of emotions, moods, personality traits and interpersonal attitudes in communication.

Several attempts have been done to integrate emotional intelligence into user interfaces, that is the ability of computers to recognize the user emotions and adapt their behavior accordingly [22], [23], [24]. Research on embodied conversational agents widely demonstrates how the ability of synthetic characters to deal with emotion recognition and expression is now considered as a key element for their believability [25], [26]. Studies have been performed to extract affective cues from language, with the goal of designing emotionally and socially intelligent interfaces to communicate either with robots [27] or software agents [28]. With respect to natural language interfaces, negative affective states, such as boredom and frustration, have been widely investigated, due to their detrimental effects on the user experience and satisfaction. It is the case of Litman and Forbes [29], who defined an annotation scheme to label emotions in tutoring dialogs along a linear scale (negative, neutral, positive) and Batliner et al. [30], who defined a method for automatically detecting emotionally critical phases in a dialogue with customers of an automatic call-center.

2.2.2 Affect detection and sentiment analysis from text

In computational linguistics, automatic detection of affective states from text is becoming increasingly important from an applicative point of view [31]. In fact, sensing emotions and other affective states from text is becoming a fundamental issue in several domains such as human-computer interaction (see, for example [22], [23], [32]) or opinion mining [33]. Statistical language learning techniques have been widely applied to such tasks, also to detect personality traits [34], [35].

Several other approaches have also been investigated. Liu et al. [1] propose a method based on large-scale real-world knowledge about the way people usually make appraisals of everyday situations. The approach exploits generic knowledge basis of commonsense to identify the six Ekman basic emotional states (happy, sad, angry, fearful, disgusted, and surprised) through text analysis. Neviarouskaya et al. [36] use a rule-based approach that includes consideration of the deep syntactic structures for emotion computation in text. Experiments are performed on different corpora to identify nine emotional labels (plus the 'neutral' one). Compared with other state-of-the-art techniques, the method shows promising results in fine-grained emotion recognition. Le Tallec et al. [27] studied how to classify emotions in speech by considering the language of hospitalized children interacting with companion robots. Only linguistic clues are considered, achieving good results in detecting emotional valence of utterances.

As far as opinion mining is concerned, sentiment analysis and the recognition of the semantic orientation of texts is also an extremely active research area (see, for example, [37], [38], [39], [40], [41], [42]). Sentiment analysis techniques may be applied to detect affective states conveyed by a text, to explore opinion with market analysis goals, to automatically analyze large collection of posts providing feedback or reviews.

Regardless of their specific application domain, the maturity reached by the techniques used in emotion and opinion detection from text suggested us the possibility to consider affect analysis as an additional source of information for enriching the set of features and hence the performance of our DA classifier.

3 Unsupervised DA Recognition

A dialogue act can be identified with the communicative goal of a given utterance, i.e. it represents its meaning at the level of illocutionary force [2]. Researchers use different labels and definitions to address this concept: Searle [3] talks about *speech act*; Schegloff [43] and Sacks [44] refer to it as *adjacency pair part*; Power [45] adopts the definition of *game move*; Cohen and Levesque [46] provide a definition of speech acts by focusing on their role in interagent communication.

Traditionally, the NLP community has employed DA annotation approaches with the drawback of being domain oriented. Only recently some efforts have been made towards unification of DA annotation [47]. In this study we refer to DAMSL (Dialogue Act Markup in Several Layers) a domain-independent annotation framework [6]. DA annotation is out of the scope of the present study hence we used already annotated data. In particular, the Switchboard employs the SWBD-DAMSL revision DAMSL [48]. Table 1 shows our set of labels: it maintains the DAMSL main peculiarity of being domain-independent and the semantics of the SWBD-DAMSL labels used for the original Switchboard annotation. The original annotation has been automatically converted in our set of tags, as reported in [5].

3.1 DA recognition: experimental setup and results

We run our experiments on the Switchboard corpus of English task-free telephone conversations [15], which involve couples of randomly selected strangers talking informally about general interest topics. Complete transcripts are distributed by the Linguistic Data Consortium. A part of them is annotated using DA labels (1155 dialogues, 205,000 utterances, 1.4 million words overall)[2]. The results presented here are obtained by randomly splitting the Switchboard in two 80/20 train/test partitions. We compare the performance of our method with the one obtained using the same partitions in a supervised framework. In particular, we used the Support Vector Machine (SVM) [49], which is regarded as a state-of-the-art technique.

Schematically, our unsupervised methodology is: (i) building a semantic similarity space in which words, set of words, text fragments can be represented homogeneously, (ii) finding seeds (words) that properly represent dialogue acts and considering their representations in the similarity space, and (iii) checking the similarity of the utterances.

To get a similarity space with the required characteristics, we used Latent Semantic Analysis (LSA), a corpus-based measure of semantic similarity [50]. In LSA, term co-occurrences in a corpus are captured by means of a dimensionality reduction operated by a singular value decomposition (SVD) on the term-by-document matrix $\mathbf{T}$ representing the corpus. The power of the model comes from the optimal dimensionality reduction [51] and choosing the best rank r' is a complex and still open problem. Empirically, it has been shown that NLP applications benefits from setting r' in the range [50,400]. For our experiments, we employ $r' = 400$.

For representing a word set or a sentence in the LSA space we use the *pseudo-document* representation technique, as described by Berry [52]. In practice, each text segment is represented in the LSA space by summing up the normalized LSA vectors of all the constituent words, using also a *tf.idf* weighting scheme [53].

For each DA we defined a set of seeds (words) representing the illocutionary force of DAs. The methodology is unsupervised[3] as we do not exploit any training material. Moreover, the LSA space is built without considering any task-specific requirement or feature. In this sense, our approach shares the inspiration of the one proposed by Collobert et al. [54], who deal with natural language processing tasks 'from scratch' using neural networks for learning task-independent language models. Table 2 shows the complete sets of seeds for each DA. Grounding on speech act theory [2], we assume the surface realization of a dialogue utterance as mainly affected by its illocutionary force. Therefore, we associate to each DA a lexical profile defined as a set of seeds (lexical cues, such as words or linguistic markers) that 'capture' the semantic of each DA label. In other words, we are trying to capture the rules underlying communication in a way that is independent from the specific conventions used in surface realization with respect to the language and

2. ftp.ldc.upenn.edu/pub/ldc/public_data/swb1_dialogact_annot.tar.gz
3. Or minimally supervised, since providing hand-specified seeds can be regarded as a minimal sort of supervision.

Label	Description	Example	% (Total Items)
INFO-REQUEST	Utterances that are pragmatically, semantically, and syntactically questions	*'What did you do when your kids were growing up?'*	7% (9189)
STATEMENT	Descriptive, narrative, personal statements	*'I usually eat a lot of fruit'*	57% (74821)
S-OPINION	Directed opinion statements	*'I think he deserves it.'*	20% (26253)
AGREE-ACCEPT	Acceptance of a proposal, plan or opinion	*'That's right'*	9% (11814)
REJECT	Disagreement with a proposal, plan, or opinion	*'I'm sorry no'*	.3% (394)
OPENING	Dialogue opening or self-introduction	*'Hello, my name is Imma'*	.2% (263)
CLOSING	Dialogue closing (e.g. farewell and wishes)	*'It's been nice talking to you.'*	2% (2625)
KIND-ATT	Kind attitude (e.g. thanking and apology)	*'Thank you very much.'*	.1% (131)
GEN-ANS	Generic answers to an Info-Request	*'Yes', 'No', 'I don't know'*	4% (5251)
total cases			131,265

TABLE 1
The set of labels employed for Dialogue Acts and their distribution in the corpus.

the application domain. Our assumption is strongly supported by the success of the natural language processing techniques for automatic dialogue act annotation using textual features (see Section 2.1). As a consequence, the seeds are general and language-independent: they are defined intuitively by two human experts (native speakers), by considering only the communicative goal and the specific semantics of each dialogue act (i.e. apologizing and politeness expression are associated to behabitive acts as KIND-ATT label in our schema), just avoiding the overlapping between seed groups as much as possible. Since our aim is to design an approach that is as general as possible, we do not consider domain words that could make easier the classification.

To assign a DA label to new utterances, we start from the sets of seeds representing the dialogue acts and we build the corresponding vectors in the LSA space. Then we compare the utterances and each DA representation in order to find the communicative act with the highest cosine similarity. To allow comparison with SVM, the performance is measured on the same test set partition used in the supervised experiment.

We reduce data sparseness using a POS-tagger and a morphological analyzer [55] to replace tokens with lemmata in the format *lemma#POS*, with no further feature selection, in both experimental settings. In addition, we augment the features of each sentence with a set of linguistic markers, defined according to the semantic of the DA labels. The addition of these markers is performed automatically, by just exploiting the output of the POS-tagger and of the morphological analyzer, according to the following rules: (i) *Wh-Qtn*, used whenever an interrogative determiner (e.g. 'what') is found; (ii) *Ask-If*, used whenever an utterance presents the pattern of a 'Yes/No' question; (iii) *I-Pers*, used for all declarative utterances whenever a verb is in the first person form, singular or plural; (iv) *Cond*, for conditional form is detected; (v) *Super*, for superlative adjectives; (vi) *Agr-Ex*, used whenever an agreement expression (e.g. 'You're right', 'I agree') is detected; (vii) *Name*, used whenever a proper name follows a self-introduction expression (e.g. 'My name is'); (viii) *Or-Clause*, used for or-clauses, that is utterance starting by 'or'. These linguistic markers were included in a DA set of seeds if found relevant for that specific dialogue act by the human experts responsible of the definition of the DA linguistic profiles (see Table 2). For example, Wh-Qtn and Ask-if are assumed to be peculiar of INFO-REQUEST while Agr-Ex, I-Pers and Name are assumed as helpful for characterizing AGREE-ACCEPT, STATEMENT and OPENING, respectively.

We evaluated the performance in terms of precision, recall and F1-measure according to the DA labels given by annotators. As a baseline we consider the most frequent label assignment for the supervised experiment (57% that is the percentage of statements in the corpus) and random DA selection for the unsupervised one (11%). We got .77 of F1 in the supervised condition, and .68 for the unsupervised one. Both results are noticeably above the baselines and are comparable to the state of the art (see section 2.1). This is particularly encouraging, especially considering that we focus only on written text (see Table 3 for the DA recognition performance on the Switchboard corpus) and that we simply consider the lexicon in utterances, without performing any word sense disambiguation nor considering deep syntactic structure of sentences.

Consistently with our goal of defining a general method for DA annotation, we compared the performance on the Switchboard corpus with the results on an Italian corpus of human-computer interactions [32]. The seeds were validated in both cases by native speakers and are the same for both languages, which is coherent with our goal of defining a language-independent method. We obtain similar performance on the two experiment (we got .66 of F1 for the Italian), confirming the independence of the approach from the two language used (see [5] for detailed results and discussion).

3.2 Error analysis

The main cause of error is the misclassification of many utterances as STATEMENT (see Table 4): statements are usually longer (in terms of number of words per utterance) and constitute the more frequent class in our corpus (57%). Hence, it is highly likely that they contain occurrences of lexical features that characterize other DAs as well and this surely affect

Label	Seeds for the original DA annotation experiment
INFO-REQ	Question_mark, interrogative determiners (*Wh-Qtn*), *Ask-If*
STATEMENT	First person verbs and pronouns (*I-Pers*)
S-OPINION	Verbs which directly express opinion or evaluation (guess, think, suppose, affect)
AGREE-ACC	yep, yeah, absolutely, correct, *Agr-Ex*
REJECT	Verbs which directly express disagreement (disagree, refute)
OPENING	Expressions of greetings (hi, hello), words and markers related to self-introduction formula, *Name*
CLOSING	Interjections/exclamations ending discourse (alright, okey, Exclamation_mark), Expressions of thanking (thank) and farewell (bye, bye-bye, goodnight, goodbye)
KIND-ATT	Lexicon which directly expresses wishes (wish), apologies (apologize), thanking (thank) and sorry-for (sorry, excuse)
GEN-ANS	no, yes

TABLE 2
The complete sets of seeds

	SVM			LSA		
Label	prec	rec	F1	prec	rec	F1
INFO-REQ	.92	.84	.88	.93	.70	.80
STATEMENT	.79	.92	.85	.70	.95	.81
S-OPINION	.66	.44	.53	.41	.07	.12
AGREE-ACC	.69	.74	.71	.68	.63	.65
REJECT	-	-	-	.01	.01	.01
OPENING	.96	.55	.70	.20	.43	.27
CLOSING	.83	.59	.69	.76	.34	.47
KIND-ATT	.85	.34	.49	.09	.47	.15
GEN-ANS	.56	.25	.35	.54	.33	.41
micro	.77	.77	.77	.68	.68	.68

TABLE 3
Evaluation of DA recognition in the two frameworks

	LSA Prediction								
Label	IR	ST	SO	AA	RJ	OP	CL	KA	GA
INFO-REQ	.70	.27	.03	-	-	-	-	-	-
STATEMENT	-	.95	.03	-	-	-	-	-	-
S-OPINION	-	.91	.07	.01	.01	-	-	-	-
AGR-ACCEPT	-	.19	.04	.63	-	-	-	-	.12
REJECT	-	.32	-	.01	.01	-	-	-	.60
OPENING	.33	.19	-	-	-	.43	-	-	-
CLOSING	.01	.47	-	.08	-	-	.34	.04	.02
KIND-ATT	-	.41	.03	-	-	-	.09	.47	-
GEN-ANS	-	.20	.01	.45	-	-	-	-	.33

TABLE 4
Confusion matrix for the LSA unsupervised classifier

	S-OPINION		**STATEMENT**		
	adjectives		*pronouns*		
	obstinate	.67	I-PERS	.60	
	overloaded	.65	I	.54	
	pathetic	.53	*adjectives*		
Expressing	satisfying	.50	nonstop	.48	*Expressing*
subjective	dirty	.50	outboard	.48	*objective*
evaluation	ridiculous	.47	bohemian	.48	*properties*
	sad	.45	inboard	.48	
	scary	.41	powdered	.48	
	horrid	.39	spiky	.47	
	fabulous	.39	salvageable	.38	
	wrongful	.38	cooperative	.35	
	terrible	.37			
	outrageous	.36			
	verbs		*verbs*		
Expressing	disqualify	.63	tarnish	.40	
attitudes	surmise	.51	have	.37	
feelings	hurt	.39	*proper nouns*		
and					
subjective	suppose	.34	of person	.48	
evaluations	preoccupy	.33	of places	.47	
	think	.31	*numbers*	.48	
	frighten	.31			
	nouns		*nouns*		
Related	indifference	.53	jumbo	.48	*Nouns*
to					
attitudes	constancy	.52	milliliter	.48	*denoting*
and affect	loyalty	.31	rhapsody	.48	*objects*
	darling	.39	pearl	.42	
	shame	.38	velvet	.37	
	danger	.27	soybean	.36	
	bastard	.27	lime	.35	
	cruelty	.27	cleanup	.35	

TABLE 5
The lexical similarity for S-OPINION and STATEMENT

the building of the LSA space. This is particularly true for S-OPINION and KIND-ATTITUDE, which are mostly misclassified as STATEMENT: the only significative difference between the two labels and the statements seems to be the wider usage of affectively loaded lexicon when conveying an opinion or expressing politeness. Recognition of such cases could be improved by enriching the data preprocessing, e.g. by exploiting information about affective lexicon.

A minor source of confounding is the misclassification of the OPENING as INFO-REQUEST. The reason is not clear yet, since the misclassified openings are not question-like in their structure. Moreover CLOSING are confounded with AGREE-ACCEPTs because of the presence of expressions like *ok*, whose common role is to express agreement but in the case of closings often introduce the farewell expressions (e.g. A: 'Well, it was nice talking to you' B: 'Ok, bye-bye'). Eventually, there is some confusion among the backchannel labels (GEN-ANS, AGREE-ACC and REJECT) due to the inherent ambiguity of common words like *yes*, *no*, and *ok*, which had been already highlighted by previous research [20].

To gain a better insight on main factor lowering the performance, that is the misclassification of OPINION as STATEMENT, we performed a study to evaluate whether an actual similarity exists in the LSA space between utterances labeled as opinions and affective lexicon.

As said before, LSA enables us to build a semantic similarity space in which words, set of words, text fragments and hence DA labels can be homogeneously represented and compared. To evaluate semantic similarity of DA labels with the lexicon in our corpus, we need to include them as pseudowords, following the methodology explained in Section 3.1 The pseudo-document representation technique is then employed to evaluate the similarity between DA labels and the words in the Switchboard. Results are summarized in Table 5 and actually show that S-OPINIONs are more similar to slanted adjectives with a non-neutral a priori polarity while STATEMENT are shown to be similar to nouns or adjectives which do not directly refer to attitudes or evaluations.

Identification of positive or negative opinions expressed linguistically is usually addressed, in the literature, in terms of sentiment analysis. At present, most work in this field was developed on monologs, such as reviews (see, e.g. [33]). Though, extending these methods to the analysis of single sentences or brief dialogue turns is not immediate even if at a first glance, sentiment analysis should work well also in these cases. The similarity study presented here is a first step towards the attempt of embedding affect analysis in order to better disambiguate dialogue acts, which is the long-term goal of our ongoing research.

4 The study

4.1 Studying the lexicon of Dialogue Acts

To address RQ1, we performed a qualitative analysis for investigating:

(a) the relationship between affective loaded lexicon of a given utterance and the communicative intention it conveys (i.e. the DA);
(b) the objectivity score of the lexicon of each DA;
(c) the salience of word categories for each DA.

To ensure the generality of our analysis, we used three lexical resources that have been developed independently and for supporting different research purposes.

In particular, to analyze the affective load of the DAs we exploit the WordNet Affect [56] lexicon to represent emotions in a LSA space. The objectivity score of each speech act is evaluated using SentiWordNet [57], whose lexicon is annotated in terms of valence (positive vs. negative) and subjectivity, for supporting general opinion mining applications. Finally, the study on word salience is based on the Linguistic Inquiry and Word Count taxonomy (LIWC), developed for supporting psycholinguistic research [58].

We are aware of the importance played by deep [36], [27] and shallow [59], [60] syntactic features in affect recognition from texts. Though, we would like to remind the reader that the long-term goal of our investigation is to better understand what are the distinctive lexical features of each DA, so as to improve the performance of our unsupervised DA classifier. Hence, we designed the three experiments, reported in the following, consistently with our approach based on lexical semantics, as described in the previous section.

4.1.1 Affective load of dialogue acts

We calculate the affective load of each DA label using the methodology described in [61]. The idea underlying the method is the distinction between *direct* and *indirect* affective words [62]. According to Ortony et al. [63], in fact, it is possible to distinguish between words that directly refer to emotional states (e.g. 'fear', 'joy', 'cheerful', 'sad') and those having only an indirect reference to an emotional state, depending on the context (e.g. the words which indicates emotional causes such as 'killer' or 'monster' or emotional responses to an event such as 'cry' or 'laugh'). For direct affective words, authors refer to the WordNet Affect [56] lexicon, which is exploited to represent emotions in an LSA space acquired from the British National Corpus (BNC)[4]. As far as indirect affective words are concerned, their affective load is evaluated by exploiting their similarity with each emotion label in the LSA space.

A-label	Example of Synsets
EMOTION	noun "anger", verb "fear"
MOOD	n. "animosity", adj. "amiable"
TRAIT	n. "aggressiveness", adj. "competitive"
COGNITIVE State	n. "confusion", adj "dazed"
PHYSICAL State	n. "illness", adj "all in"
HEDONIC SIGNAL	n. "hurt", n. "suffering"
Emot.-Elicit. SITUATION	n. "awkwardness"
Emot. RESPONSE	n. "cold sweat", v. "tremble"
BEHAVIOUR	n. "offense", adj. "inhibited"
ATTITUDE	n. "intolerance", n. "defensive"
SENSATION	n. "coldness", v. "feel"

TABLE 6
A-labels in WordNet Affect with examples

WordNet Affect[5] [56] is an extension of the WordNet database [65], which employs affective labels (*a-labels*) to annotate the WordNet synsets. One or more a-labels may be assigned to a synset. The resource includes also a-labels representing moods, situations eliciting emotions or emotional responses (see examples in Table 6). For the purpose of this study, we considered six basic emotion labels (anger, digust, fear, joy, sadness, surprise): starting with WordNet Affect, six lists of affective words are collected, according

4. The British National Corpus (BNC) is s balanced text corpus [64]. It consists of a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of modern English, both spoken and written. The written part of the BNC includes, for example, extracts from newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, school and university essays, among many other kinds of text. The spoken part consists of orthographic transcriptions of unscripted informal conversations (recorded by volunteers selected from different age, region and social classes in a demographically balanced way) and spoken language collected in different contexts, ranging from formal business or government meetings to radio shows and phone-ins.

5. This resource is freely available for research purposes at http://wndomains.fbk.eu

to the approach and the features of the resources employed in [61]. LSA is then used to learn, in an unsupervised setting, a vector space from the BNC. As said before, LSA has the advantage of allowing homogeneous representation and comparison of words, text fragments or entire documents, using the pseudo-document technique. Hence, we are able to evaluate the semantic similarity among generic terms and affective lexical concept: a synset in WordNet, as well as all the words labeled as carrying a particular emotion, can be represented, in the LSA space, by performing the pseudo-document technique on all the words contained in the synset.

Among the various way in which an emotion may be represented in the LSA space [61], we choose to represent each emotion label as the vector representing the synset of the emotion, i.e. in addition to the word denoting an emotion (e.g. 'anger'), its synonyms from the respective WordNet synset are also used (e.g. 'anger', 'choler', 'ire'). We compute the affective load of a given textual input in terms of its similarity with the vector representing the emotion. This can be done by computing the similarity among the generic terms (including the indirect affective words) in the input text and the affective categories. Hence, the affective load of a given utterance is calculated in terms of its lexical similarity with respect to each of the six emotion labels, obtaining six scores of similarity for each utterance (one per emotion). Each score is normalized by the numbers of utterances receiving the same DA label in the corpus. The overall affective load of the sentence is then calculated as the average of its similarity scores with each emotion label. The overall affective load of each DA label is then calculated by the average score observed per each subset of utterances in the corpus receiving the same label (see Table 7).

Results are quite encouraging and suggested us to further investigate the relationship that exists between the communicative goal of an utterance and the use of affective loaded lexicon. The overall affective load of the corpus (.1182) is calculated as the average of the affective load scores of all utterances (in this case, the scores are normalized by the numbers of all utterances in the Switchboard, regardless of their speech act label). It is interesting to observe how this overall score reflect the score of INFO-REQUEST, which is reasonable to assume as neutral in terms of emotional content. S-OPINION is the DA with the highest affective load, immediately followed by KIN-DATT due to the high frequency of politeness expressions in such utterances (see Table 8 for examples), which is consistent with the error analysis and similarity study performed in Section 3.2.

Label	Affective Load	St. Dev
S-OPINION	.1439	.03
KIND-ATT	.1411	.07
STATEMENT	.1300	.04
INFO-REQ	.1142	.04
CLOSING	.0671	.05
REJECT	.0644	.06
OPENING	.0439	.05
AGREE-ACC	.0408	.06
GEN-ANS	.0331	.05
Overall Affective Load of Switchboard		.1182

TABLE 7
Affective load of DA labels

S-OPINION
Gosh uh, it's getting *pathetic* now, absolutely *pathetic*.
They're just *horrid*, you'll have nightmares, you know.
That's no way to make a decision on some *terrible* problem.
They are just gems of shows. Really, *fabulous* in every way.
And, oh, that is so good. *Delicious*.
KIND-ATTITUDE
I'm *sorry*, I really feel strongly about this.
Sorry, now I'm probably going to upset you.
I *hate* to do it on this call.

TABLE 8
Slanted lexicon in S-OPINION and KIND-ATT

4.1.2 Objectivity score of dialogue acts

In search for further support to the findings of the analysis of affective load of DA, in this section we investigate the subjectivity of each DA label using SentiWordNet 3.0 [57].

SentiWordNet is based on the WordNet lexicon and associates each synset of WordNet 2.0 to *Pos*, *Neg* and *Obj* numerical scores describing, respectively, how positive, negative and objective are the terms contained in a sysnset. The scores range from 0.0 to 1.0 and their sum is 1.0 for each synset (the objectivity score is calculated as Obj = 1-(Pos-Neg)). We use SentiWordNet because it provides graded scores that enable capturing positivity/negativity nuances of each synset. Moreover, the semi-automatic annotation of the lexicon has been performed so as to ensure generality and domain-independence of the resource.

We calculate the average objectivity score of each DA, that is the normalized average of the positivity/negativity score of each utterance by considering all its verbs, adverbs, nouns and adjectives. The SentiWordNet scores are provided for each sense of WordNet. Though, in our DA classification approach we do not perform word-sense disambiguation. On the contrary, we have a preliminary part-of-speech tagging phase, during the data preprocessing. Therefore, to be consistent with our DA classification approach, we decided to choose the most frequent synset for each word (i.e. the first sense of WordNet), with respect to the recognized part of speech. The most frequent sense of a word, in fact, is ususally considered as a baseline for word sense disambiguation tasks. Moreover we decide to avoid too general terms by considering only the words with low polisemy (i.e. lemmata with less than 5 senses), calculated on WordNet.

For each utterance of the corpus, the positive, negative and objective scores are evaluated and then the

DA Label	Average SentiWordNet Scores			
	Positive	Negative	Objective	Average Polisemy
KIND-ATT	0.05	0.23	0.72	5.53
S-OPINION	0.08	0.05	0.88	10.03
STATEMENT	0.06	0.04	0.91	10.17
INFO-REQ	0.03	0.02	0.95	10.14
CLOSING	0.03	0.02	0.95	5.77
AGREE-ACC	0.03	0.01	0.97	5.04
REJECT	0.02	0.01	0.97	4.25
GEN-ANS	0.02	0.01	0.98	2.62
OPENING	0.01	0.00	0.99	4.89

TABLE 9
Subjectivity of DA labels

average scores for each DA are calculated by grouping all utterances according to their label. Results are shown in Table 9 and it is interesting to see how they are consistent with those obtained for the affective load study described in the previous section. In fact, utterances expressing opinions and politeness expression (S-OPINION and KIND-ATT) receive again the lower objectivity score.

4.1.3 Identifying dominant lexical categories in DA

To better understand what are the most distinctive lexical features (i.e. word classes, with particular focus on categories including affective lexicon) for each DA, we performed a qualitative investigation of the lexicon in the Switchboard corpus. We followed the methodology described in [66] to calculate a score associated with a given class of words, in order to evaluate the relevance of each class with respect to a specific DA.

Let C be a class of words $C = W_1, W_2, ..., W_n$ and *da* a generic dialogue act, belonging to the set employed for this study (see Table 1). We can build the corpus DA including all utterances in our data set that have been labeled as da (e.g. the complete set of all INFO-REQUEST), as well as the complementary corpus $\neg DA$, which includes all the utterances annotated differently. We compute the *dominance score* for the class C in the generic dialogue act DA as

$$Dominance_{DA}(C) = \frac{Coverage_{DA}(C)}{Coverage_{\neg DA}(C)} \quad (1)$$

The class coverage for the DA is calculated as

$$Coverage_{DA}(C) = \frac{\sum_{W_i \in C} Frequency_{DA}(W_i)}{Size_{DA}}$$

where $Frequency_{DA}(W_i)$ is the total number of occurrences of all words in C in DA and $Size_{DA}$ is the dimension of DA in words. Analogously, the class coverage for the rest of the corpus $\neg DA$ is calculated as

$$Coverage_{\neg DA}(C) = \frac{\sum_{W_i \in C} Frequency_{\neg DA}(W_i)}{Size_{\neg DA}}$$

A dominance score close to 1 indicates that C has a similar distribution for both DA and the rest of the corpus (that is, C is not salient for da). On the contrary, a score significantly higher than 1 indicates a high salience of a class of words for a given DA.

In our study, we refer to the word classes defined in the Linguistic Inquiry and Word Count (LIWC) taxonomy, developed in the scope of psycholinguistic research [58]. We do not consider domain specific categories of words (e.g. School, Money, Leisure etc.) in order to make the analysis consistent with our goal of defining a domain-independent approach for DA annotation.

The LIWC taxonomy organizes words into psychologically meaningful categories and have been used for a wide range of psycholinguistics experimental settings, including investigation on emotions, social relationships, thinking styles, and so on [8]. LIWC is organized according to the assumption that words and language reflect most part of cognitive and emotional phenomena involved in communication. Language is seen as a medium by which 'cognitive, personality, clinical and social psychologists attempt to understand human beings' [8]. The LIWC categories were initially developed to draw distinctions between negative and positive emotion words and subsequently expanded to contemplate also thinking style (e.g. causal reflection) and other psycholinguistic phenomena. At present, LIWC includes 80 categories, along several language dimensions: some of them are objective, such as the 'Articles' category, while other could not be so straightforward. It is the case of subjective categories (such as AFFECT, NEGEMO and POSEMO) for which a dictionary was created by starting from candidate word lists which were then revised and/or extended according to human judges' ratings. Table 11 shows some of the categories we consider in our study with sample words.

LIWC has been successfully exploited by a wide range of psycholinguistic studies (see [8] for an overview), demonstrating how language analysis can provide evidence on people's attentional focus. In particular, the use of pronouns informs about subject's attention: people setting in front of mirrors tend to focus on self hence using more first person pronouns to indicate self-reference (words like 'I' and 'we') as well as people involved in positive political ads; negative ads, on the contrary, show a prevalence of reference to others (words like 'he' or 'they'). In the case of dialogue acts we could hypothesize that people expressing feelings and making statements probably use more self-reference with respect to people who ask question, whose attentional focus is on the interlocutor (assumption confirmed by the results, as shown in Table 10).

Analogously, the tense of verbs may be an indicator of temporal focus and may serve as an indicator of intentions: e.g. the imperative mood may be used for giving an order while the past tense may be employed in telling stories or referring facts or describing the world, typically in statements.

Moreover, the language may be an indicator of emotionality: language emotionality, though, should be seen not only as a directly expression of affective states, since it may extend beyond the simple expression of emotions or feelings and relate to other key language elements, as discussed earlier.

Also, natural language can provide information on how people process information. For example, the use of causal words (e.g. 'because', 'affect', 'hence') and insight words ('think', 'know', 'consider') may indicate a cognitive activity involving appraisal mechanism and could be of potential interest for improving the recognition of opinions.

Language may also serve as an indicator of social relationships between participants to the dialogue, as a cue of the speaker's honesty and deception or to draw differences between individuals with respect to personality traits, age and sex.

To conclude, the use of language is a very powerful indicator of what people's attentional focus, intentions, emotional states, personality and motivations and LIWC represents a particularly interesting resource in this sense. Hence we decide to exploit the LIWC taxonomy to perform our qualitative study about salience of the word classes in dialogue acts. Excluding from this investigation the domain-related categories will ensure, once again, the domain-independence of our approach and the general validity of our findings.

Table 10 shows the ranking for the most salient word classes for each DA with their dominance score. Sample words for each class are provided in Table 11.

Opinion		Statement		Kind-Att	
FUTURE	2.00	PAST	2.17	NEGEMO	19.14
NEGEMO	1.85	I,SELF,WE	2	AFFECT	7.95
SAD	1.69	INCL	1.41	POSEMO	5.43
INSIGHT	1.56	SEE	1.30	COMM	4.51
ANGER	1.54	MOTION	1.25	INHIB	2.68
DISCREP	1.47	HEAR	1.18	ANGER	2.61
OPTIM	1.49	SENSES	1.17	SELF, FEEL	2.3
FEEL	1.44			ANX	1.87
SWEAR	1.40				
COGMECH	1.37				
Reject		**Agree-acc**		**Opening**	
NEGATE	14.54	ASSENT	75.32	COMM	27.65
METAPH	1.91	CERTAIN	4.64	ASSENT	3.22
NEGEMO	1.60	POSEMO	2.67	SOCIAL	3.10
INHIB	1.22	AFFECT	2.22	CAUSE	3.02
		OPTIM	2.12	HEAR	2.10
Closing		**Info-Req**		**Gen-Ans**	
HEAR	8.10	YOU	3.73	ASSENT	38.21
ASSENT	6.75	CAUSE	1.88	NEGATE	7.15
COMM	6.42	OTHREF	1.73		

TABLE 10
Dominant word classes for each DA with their scores

4.2 Affect Analysis and DA identification

In this Section, we address RQ2, that is whether it is possible to improve the DA classification by exploiting affective lexicon. In particular, we exploit the insights derived from the word salience experiment

Class	Sample words
ACHIEVE	accomplish, award, beaten, ahed, celebrating
AFFECT	wrong, warm, vulnerable, violent, unpleasant
ANGER	molest, offend, outrage, revenge, ridicule
ANX	worried, terrifying, stress, scare, hesitate
ASSENT	accept, alright, fine, yep, yeah
CAUSE	affect, basis, because, coz, depends, infers,
CERTAIN	always, all, very, truly, completely, totally
COGMECH	acknowledge, admit, become, believe
COMM	call, chat, confess, describe, discuss, e-mail
DISCREP	but, expect, hope, if, must, need, should, wishing
EXCL	although, besides, except, but
FEEL	tries, senses, pain, hold, grab, feel
FUTURE	be, I'll, may, might, will, won't, you'll
HEAR	ask, call, discuss, ear, listen, say, sound
I	I, myself, mine
INCL	also, altogether, and, here, plus
INHIB	block, constraint, control, forbid, limit, prevent
INSIGHT	believe, think, know, see, understand, found
METAPH	god, die, sacred, mercy, sin, dead, hell, soul
MOTION	action,bring, carry, cross, deliver, drive, enter
NEGATE	aren't, don't, neither, no, never, zero
NEGEMO	abandon, anger, boring, cry, danger, depressed
OPTIM	best, ready, hope, accepts, determined, won, super
OTHER	she, her, they, his, them, him
OTHREF	anybody, anyone, everybody, he'll, he's, our
PAST	accepted, ago, became, called, did, guessed
POSEMO	won, wealth, triumph, treasure, wisdom
POSFEEL	sentimental, romantic, passion, love, liking,
PRESENT	accept, begin, believe, carry, happen, have,
PRONOUN	us, we, you, thou, somebody, she, our
SAD	alone, cry, depressed, hopeless, miss, pity, useless
SEE	appear, show, see, eye, look, vision, watch, witness
SELF	our, myself, mine, ours
SENSES	witness, touch, tell, talk, look, listen, say, read
SIMILES	like
SOCIAL	ya, ye, you, you'd, you'll, your
SWEAR	damn, crap, hell, bastard
TENTAT	alot, any, anywhere, bet, depend, hope, lucky
WE	us, we, our, ourselves
YOU	you, thou

TABLE 11
LIWC word classes with sample words

(see Section 4.1.3). The findings of Sections 4.1.1 and 4.1.2 suggest that a relationship exists between the use of either affectively loaded and subjective lexicon and the surface realization of specific speech acts. Though, we would like to remind the reader that one of the goal of this ongoing research is to incorporate affect analysis in our unsupervised method for DA annotation. Consistently with our approach based on lexical semantics, we selected LIWC as a resource to be included in this experiment over WordNet Affect and SentiWordnet because it immediately fits in the approach for DA recognition described in Section 3.1.

We augment the features of each sentence in our corpus with a set of linguistic markers, corresponding to the labels of the word classes in the LIWC taxonomy. According to the evidences reported in Section 4.1.3, we argue that these features could play an important role in defining the linguistic profile of each DA. The addition of these markers is performed automatically, by just exploiting the output of the POS-tagger and of the morphological analyzer. The

natural language input:
(a) 'I just don't care.'
correspondent dataset item:
(a) .#PUN:1 do#v:1 i:1 that:1 I_PERS:1 not:1 just#adv:1 care#v:1 PRONOUN:1 POSFEEL:1 NEGATE:1 EXCL:1 PRESENT:2 SELF:1 POSEMO:1 I:1 AFFECT:1 TENTAT:1

TABLE 12
Enriching utterances with LIWC categories

	SVM with LIWC			SVM (prev. setting)		
Label	prec	rec	F1			
INFO-REQ	.92	.84	.88	.92	.84	.88
STATEMENT	.79	.93	.86	.79	.92	.85
S-OPINION	.69	.44	.54	.66	.44	.53
AGREE-ACC	.67	.77	.72	.69	.74	.71
REJECT	-	-	-	-	-	-
OPENING	.96	.58	.72	.96	.55	.70
CLOSING	.84	.56	.68	.83	.59	.69
KIND-ATT	1.0	.13	.22	.85	.34	.49
GEN-ANS	.60	.19	.29	.56	.25	.35
micro	.77	.77	.77	.77	.77	.77

TABLE 13
Enriching the corpus with features based on LIWC

text of each utterance is enriched by adding the class labels whenever an occurrence of a word belonging to a specific class is found, as shown in the example in Table 12, consistently with the approach adopted for the linguistic markers described in Section 3.1. To remain consistent with our approach, no disambiguation is performed and we do not take into account the deep syntactic structure of sentences nor context information. We run SVM on the same train/test partitions and procedure described in Section 3.1. Results are reported in Table 13.

4.3 Discussion

In this section we have addressed our research questions. In particular, in Section 4.1 we describe a qualitative investigation of the lexicon of dialogue acts aimed at verifying the relation between the affective lexicon and the communicative goal of utterances in a dialogue (RQ1). Then, in Section 4.2 we verify what is the role played by affective language in dialogue act identification (RQ2), by exploiting the LIWC word classes to enrich the features of our corpus.

To address RQ1 we have performed a qualitative analysis using three different resources and approaches, developed in previous research. The results of the experiments on affective load and objectivity score of DA labels (see sections 4.1.1 and 4.1.2, respectively) suggest that a relation exists between affective lexicon and the communicative intention of the speaker, at least in the Switchboard corpus. In particular, the highest affective load and objectivity scores are observed for S-OPINION and KIND-ATTITUDE labels. Results are consistent with both the similarity study (Table 5) and the analysis of word class dominance for each DA (Section 4.1.3). Specifically, the dominance study highlights a prevalence of negative emotions in the expression of opinions, while words referring to both, positive and negative affective states, are used for kind-attitude utterances. Also, the class FEEL is relevant to both labels.

The wider use of emotion lexicon for KIND-ATTITUDE is consistent with their 'expressive' nature [4]. Of course, and according to Austin's definition of 'Behabitives' [2], the fact that affective lexicon is used in the formulation of politeness expression (which is typical for KIND-ATTITUDE utterances) does not necessary mean that the speaker is reporting about an emotion actually felt while talking/writing. Still, we believe this information about the use of affective lexicon in both opinions and kind attitude expressions may be successfully exploited to improve the DA classification performance in our future research, by enriching the set of seeds for both labels using the relevant LIWC affective word classes according to their dominance score (Table 10).

Moreover, we observe how distinct lexical choices are operated by speakers when formulating STATEMENTs and S-OPINIONs. This is interesting since the confounding between these two labels is the main cause of error of our DA classifier. According to the dominance analysis, statements are mainly expressed using the past tense, the first person pronouns and expressions of inclusion (*also*, *altogether*, *plus*) while opinions mainly use the future tense. Also, when formulating statements people talk about facts, using lexicon related to physical actions (MOTION), the five senses and the perception of the world (SENSES). On the contrary, when expressing opinions people mainly report about their feelings (FEEL) and beliefs (COGMECH). This result is coherent with the descriptive/narrative nature of statements [2], [3], [67] in contrast with the subjective connotation of opinions, which are rather connected to appraisal and evaluation mechanisms. Moreover, it clearly reflects the criterion adopted in the original Switchboard annotation [48]: the statement-non-opinion tag, in fact, is used when the speaker is telling a story and the topic is personal while cues as *I think*, *I believe*, *I mean* etc. are explicitly defined as possible indicators for distinguishing opinions from objective statements.

As far as backchannel acts are concerned, we observe a clear differentiation in the lexicon used for expressing agreement and disagreement: ASSENT, CERTAIN and OPTIM categories are highly salient for the AGREE-ACCEPT label while negation (NEGATE) and exclamations (METAPH) are salient for REJECT. While being quite intuitive, this clear distinction in the formulation of (dis)agreement expression was not originally reflected in the recognition performance for the AGREE-ACCEPT and, in particular, the REJECT labels (see Tables 3 and [5] for a detailed discussion). In fact, the classes ASSENT and NEGATE are also relevant for the general answer, which confirm our previous findings about the misclassification due to ambiguous words as *right* or *yes/no* which can be seen as general answers, acknowledgement and (dis)agreement expressions. This is consistent with previous work by Jurafsky et. al [20] about inherent ambiguity of lexicon used for backchannel signaling, which appears to be the same used in most expression of agreement. Moreover, backchannel acts have been shown to be 'recalcitrant' [17] with respect to classification, even in presence of context information.

OPENING and CLOSING share the common characteristic of being both used for meta-communication goals that is, respectively, for beginning and ending the interaction. Hence, they both show linguistic features that relates to their role in the discourse, like the lexicon included in the COMM and HEAR category (e.g. verbs like *call*, *chat*, *discuss*, *talk*, etc.). For example, the category HEAR is salient for CLOSING because one of the more common ways of terminating the dialogue is to use sentences such as 'It's been nice *talking* to you'). Also, they show lexical features dependent on the interaction modality (e.g people opening by saying 'I *call* from...'), since Switchboard includes telephone conversations).

Finally, the YOU and OTHREF categories seem to be relevant for the INFO-REQUEST, which clearly indicates the attentional focus [58] on the interlocutor.

The findings derived from the dominance study are highly consistent with the indication provided by the affective load and the positive, negative and objective scores for each DA, thus providing a positive answer to RQ1 and direction for addressing RQ2. That is, they confirm that a relation exists between the use of affective lexicon and the communicative intention associated to a given speech act label, at least in the Switchboard corpus and suggest how to exploit affective word classes for improving automatic DA identification. This is particularly true for S-OPINION and KIND-ATTITUDE, which appear as the most affectively loaded labels and for which we observe a dominance of affective LIWC categories.

As far as RQ2 is concerned, though, the experiments run in Section 4.2 did not provide full evidence of the role played by the affect analysis in disambiguating affectively loaded dialogue acts. The overall performance obtained by incorporating the finding of affective analysis is the same observed for the supervised framework (micro=.77, see Table 3) and our hypothesis that LIWC allows our DA classifier to better capture the different lexical choices that people operates when formulating different dialogue acts only reflects in a slight improvement of the recognition performance for some of the DA labels.

F1 for S-OPINION utterances observe a slight improvement, mainly due to an increased precision, which is still encouraging if we consider that sentiment analysis and identification of subjectivity in text is a research domain *per se*. Moreover, we observe a noticeable increase of the precision of KIND-ATTITUDE, thanks to the exploitation of the word classes that typically occur in behabitives. Also less affectively loaded DA benefits from the exploitation of LIWC word classes in our sets of seeds. In particular, the confounding between INFO-REQUEST and OPENING is drastically reduced: the high dominance of the COMM word class for OPENING and CLOSING utterances denote they role of these speech acts in the management of the dialogue dynamics. This nature of being 'meta' communication actions makes this DA different from the others and is well captured by the use of LIWC word class labels as features in the LSA. Still, the overall performance is only slightly increased and further research is needed to address the role of affect analysis in dialogue act recognition in both supervised and unsupervised frameworks.

5 Final Remarks and Conclusions

The long-term goal of our research is to define an unsupervised approach for Dialogue Act labelling. The method has to be independent from the language, domain, size, interaction scenario of the referred corpus, focusing only on lexical analysis. In our previous work [5] some preliminary steps have been done toward the achievement of this goal. In this paper we proposed a qualitative study of the lexicon of dialogue acts in order to better understand what are the most salient and distinctive lexical features for DA profiling. In particular we investigated the relationship between the affective load of utterances and their communicative goal and exploit affect analysis for improving DA recognition.

The experimental results suggest that a relationship exists between the use of affective lexicon and the communicative intention of an utterance, at least in the Switchboard corpus of telephone conversations (see Section 4.1). People do actually make more frequent use of affective loaded lexicon when conveying certain dialogue acts (i.e. opinions and behabitives) and we have shown how state of the art techniques for affective language analysis are able to capture this phenomenon in spontaneous natural language conversations. Our findings are consistent with previous research and support the intuition on which we based our first research question RQ1, namely *if a relationship exist between the affective lexicon and the communicative goal of an utterance*. Though, in spite of the positive evidence provided to RQ1, we did not find full support to our second research question RQ2, that is *if affect analysis plays a role in DA identification*.

We are aware of the main limitations of our study and we need to replicate the experiments on different annotated corpora to verify the general validity of our findings. Moreover, we intend to improve the DA recognition by including the consideration of the affective load in our unsupervised recognition framework. In the present contribution we exploited LIWC, WordNet Affect and SentiWordnet as candidate resources. It is our plan for future research to include the consideration of other linguistic resources. Among the envisaged resources are the NRC Emotion Lexicon, a list of words and their associations with eight emotions and two sentiments (negative and positive) manually annotated through Amazon's Mechanical Turk [68] and the Subjectivity Lexicon organized as a list of subjectivity clues[6] and already successfully exploited in sentiment analysis research [69].

Moreover, it will be paramount to include the consideration of different corpora in our further replication of the present study in order to verify the generality of our findings supporting RQ1. In particular, having in mind the long-term goal of exploiting results of affect analysis from text to improve DA recognition, it would be interesting to this information to deal with the misclassification of opinions as statements, even if we are aware that this constitute a distinct research field in itself. Along this perspective, DA recognition could serve also as a basis for conversational analysis aimed at improving a fine-grained opinion mining in dialogues.

REFERENCES

[1] H. Liu, H. Lieberman, and T. Selker, "A model of textual affect sensing using real-world knowledge," in *Proceedings of the 8th international conference on Intelligent user interfaces*, ser. IUI '03. New York, NY, USA: ACM, 2003, pp. 125–132.

[2] J. Austin, *How to do Things with Words*. New York: Oxford University Press, 1962.

[3] J. Searle, *Speech Acts: An Essay in the Philosophy of Language*. Cambridge, London: Cambridge University Press, 1969.

[4] N. Guiraud, D. Longin, E. Lorini, S. Pesty, and J. Rivière, "The face of emotions: a logical formalization of expressive speech acts," in *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, ser. AAMAS '11. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2011, pp. 1031–1038.

[5] N. Novielli and C. Strapparava, "Exploring the lexical semantics of dialogue acts," *Journal of Computational Linguistics and Applications*, vol. 1, no. 1-2, pp. 9–26, 2010.

[6] M. Core and J. Allen, "Coding dialogs with the DAMSL annotation scheme," in *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, Cambridge, MA, 1997, pp. 28–35.

[7] I. Poggi and E. Magno Caldognetto, "Italian parlato," in *Il Parlato Emotivo. Aspetti Cognitivi, Linguistici e Fonetici*, 2004.

[8] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of Language and Social Psychology*, 2010.

[9] W. Bosma and E. André, "Exploiting emotions to disambiguate dialogue acts," in *IUI '04: Proceedings of the 9th international conference on Intelligent user interfaces*. New York, NY, USA: ACM, 2004, pp. 85–92.

[10] K. E. Boyer, J. F. Grafsgaard, E. Y. Ha, R. Phillips, and J. C. Lester, "An affect-enriched dialogue act classification model for task-oriented dialogue," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, ser. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1190–1199.

[11] S. Somasundaran, G. Namata, J. Wiebe, and L. Getoor, "Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, ser. EMNLP '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 170–179.

[12] K. Samuel, S. Carberry, and K. Vijay-Shanker, "Dialogue act tagging with transformation-based learning," in *Proceedings of the 17th international conference on Computational linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 1998, pp. 1150–1156.

[13] N. Reithinger and M. Klesen, "Dialogue act classification using language models," in *In Proceedings of EuroSpeech-97*, 1997, pp. 2235–2238.

6. http://mpqa.cs.pitt.edu/

[14] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. V. Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[15] J. Godfrey, E. Holliman, and J. McDaniel, "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: IEEE, 1992, pp. 517–520.

[16] S. Quarteroni, A. V. Ivanov, and G. Riccardi, "Simultaneous dialog act segmentation and classification from human-human spoken conversations," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic*. IEEE, 2011, pp. 5596–5599.

[17] A. Venkataraman, Y. Liu, E. Shriberg, and A. Stolcke, "Does active learning help automatic dialog act tagging in meeting data?" in *Proceedings of EUROSPEECH-05*, Lisbon, Portugal, 2005.

[18] V. Warnke, R. Kompe, H. Niemann, and E. Nöth, "Integrated dialog act segmentation and classification using prosodic features and language models," in *Proceedings of 5th European Conference on Speech Communication and Technology*, vol. 1, Rhodes, Greece, 1997, pp. 207–210.

[19] B. J. Grosz and C. L. Sidner, "Attention, intentions, and the structure of discourse," *Comput. Linguist.*, vol. 12, no. 3, pp. 175–204, 1986.

[20] D. Jurafsky, E. Shriberg, B. Fox, and T. Curl, "Lexical, prosodic, and syntactic cues for dialog acts," in *Proceedings of ACL/COLING 98*, Montreal, 1998, pp. 114–120.

[21] R. Picard, *Affective computing*. MIT Press, 1997.

[22] C. Conati, "Probabilistic assessment of user's emotions in educational games," *Applied Artificial Intelligence*, vol. 16, pp. 555–575, 2002.

[23] R. W. Picard and J. Klein, "Computers that recognise and respond to user emotion: Theoretical and practical implications," MIT Media Lab, Tech. Rep., 2001. [Online]. Available: http://pubs.media.mit.edu/pubs/papers/TR-538.pdf

[24] F. de Rosis, A. Batliner, N. Novielli, and S. Steidl, "'You are Sooo Cool, Valentina!' Recognizing Social Attitude in Speech-Based Dialogues with an ECA," in *Affective Computing and Intelligent Interaction*, ser. LNCS, A. Paiva, R. Prada, and R. W. Picard, Eds., Berlin-Heidelberg, 2007, pp. 179–190.

[25] T. Bickmore and J. Cassell, "Social dialogue with embodied conversational agents," in *Advances in Natural Multimodal Dialogue Systems*, J. van Kuppevelt, L. Dybkjaer, and N. O. Bernsen, Eds. Springer, 2005, pp. 23–54. [Online]. Available: http://dx.doi.org/10.1007/1-4020-3933-6_2

[26] J. Bates, "The role of emotion in believable agents," *Commun. ACM*, vol. 37, no. 7, pp. 122–125, 1994.

[27] M. Le Tallec, J.-Y. Antoine, J. Villaneau, and D. Duhaut, "Affective interaction with a companion robot for hospitalized children: a linguistically based model for emotion detection," in *Proceedings of the Fifth Language and Technology Conference (LTC2011*, 2011.

[28] N. Novielli, F. de Rosis, and I. Mazzotta, "User attitude towards an embodied conversational agent: Effects of the interaction mode," *Journal of Pragmatics*, vol. 42, no. 9, pp. 2385–2397, 2012.

[29] D. Litman, K. Forbes, and S. Silliman, "Towards emotion prediction in spoken tutoring dialogues," in *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 52–54.

[30] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "How to find trouble in communication," *Speech Commun.*, vol. 40, no. 1-2, pp. 117–143, 2003.

[31] C. Strapparava and R. Mihalcea, "SemEval-2007 task 14: Affective Text," in *Proceedings of the 4^{th} International Workshop on Semantic Evaluations (SemEval 2007)*, Prague, 2007, pp. 70–74.

[32] G. Clarizio, I. Mazzotta, N. Novielli, and F. deRosis, "Social attitude towards a conversational character," in *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication*, Hatfield, UK, 2006, pp. 2–7.

[33] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[34] A. J. Gill and J. Oberlander, "Taking care of the linguistic features of extraversion," in *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 2002.
[35] A. J. Gill, S. Nowson, and J. Oberlander, "What are they blogging about? personality, topic and motivation in blogs," in *ICWSM 2009*, 2009.
[36] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Affect analysis model: Novel rule-based approach to affect sensing from text," *Nat. Lang. Eng.*, vol. 17, no. 1, pp. 95–135, 2011.
[37] P. Turney and M. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Transactions on Information Systems*, vol. 21, pp. 315–346, 2003.
[38] B. Pang and L. Lee, "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts," in *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2004, p. 271.
[39] J. Wiebe and C. Cardie, "Annotating expressions of opinions and emotions in language. language resources and evaluation," in *Language Resources and Evaluation (formerly Computers and the Humanities*, 2005, p. 2005.
[40] R. Mihalcea, "A corpus-based approach to finding happiness," in *In AAAI 2006 Symposium on Computational Approaches to Analysing Weblogs*. AAAI Press, 2006, pp. 139–144.
[41] C. Akkaya, J. Wiebe, A. Conrad, and R. Mihalcea, "Improving the impact of subjectivity word sense disambiguation on contextual opinion analysis," in *Proc. of the Fifteenth Conference on Computational Natural Language Learning*, ser. CoNLL '11. Association for Computational Linguistics, 2011, pp. 87–96.
[42] C. Strapparava, O. Stock, and R. Mihalcea, "Computational humour," in *Emotion-Oriented Systems: The HUMAINE Handbook, Cognitive Technologies*, R. Cowie, P. Petta, and C. Pelachaud, Eds. Springer, 2011.
[43] E. Schegloff, "Sequencing in conversational openings," *American Anthropologist*, vol. 70, pp. 1075–1095, 1968.
[44] H. Sacks, E. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.
[45] R. Power, "The organisation of purposeful dialogues," *Linguistics*, vol. 17, pp. 107–152, 1979.
[46] P. R. Cohen and H. J. Levesque, "Communicative actions for artificial agents," in *Proc. of the First International Conference on Multi-Agent Systems*. AAAI Press, 1995, pp. 65–72.
[47] D. Traum, "20 questions for dialogue act taxonomies," *Journal of Semantics*, vol. 17, no. 1, pp. 7–30, 2000.
[48] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13." University of Colorado Institute of Cognitive Science, Technical Report 97-01, 1997.
[49] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.
[50] T. Landauer, P. Foltz, and D. Laham, "Introduction to latent semantic analysis," *Discourse Processes*, vol. 25, 1998.
[51] T. K. Landauer and S. T. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge," *Psychological Review*, vol. 104, no. 2, pp. 211–240, 1997.
[52] M. Berry, "Large-scale sparse singular value computations," *International Journal of Supercomputer Applications*, vol. 6, no. 1, 1992.
[53] A. Gliozzo and C. Strapparava, "Domains kernels for text categorization," in *Proc. of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, University of Michigan, Ann Arbor, 2005, pp. 56–63.
[54] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.
[55] E. Pianta, C. Girardi, and R. Zanoli, "The TextPro tool suite," in *Proceedings of LREC-08*, Marrakech, Morocco, 2008.
[56] C. Strapparava and A. Valitutti, "WordNet-Affect: an affective extension of WordNet," in *Proceedings of LREC*, vol. 4, 2004, pp. 1083–1086.
[57] A. E. S. Baccianella and F. Sebastiani, "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), 2010.
[58] J. Pennebaker and M. Francis, "Linguistic inquiry and word count: LIWC," 2001, erlbaum Publishers.
[59] L. Polanyi and A. Zaenen, "Contextual valence shifters," in *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 2004.
[60] M. Ptaszynksi, P. Dybala, W. Shi, R. Tzepka, and K. Araki, "Contextual affect analysis: A system for verification of emotion appropriateness supported with contextual valence shifters," *International Journal of Biometrics*, vol. 2, no. 2, pp. 134–154, 2010.
[61] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*. New York, NY, USA: ACM, 2008, pp. 1556–1560.
[62] C. Strapparava, A. Valitutti, and O. Stock, "The affective weight of lexicon," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, 2006.
[63] A. Ortony, C. G. L., and M. A. Foss, "Introduction to latent semantic analysis," *Journal of Personality and Social Psychology*, vol. 53, pp. 751–766, 1987.
[64] BNC-Consortium, "British national corpus." http://www.hcu.ox.ac.uk/BNC/, Humanities Computing Unit of Oxford University, 2000.
[65] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.
[66] R. Mihalcea and S. Pulman, "Linguistic ethnography: Identifying dominant word classes in text," in *Proceeding of Computational Linguistics and Intelligent Text Processing (CICLing-09)*, 2009.
[67] J. R. Searle, *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge, MA, USA: Cambridge University Press, 1979.
[68] S. Mohammad and P. Turney, "Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Association for Computational Linguistics, 2010, pp. 26–34.
[69] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05. Association for Computational Linguistics, 2005, pp. 347–354.

Nicole Novielli is a research fellow at the Department of Computer Science, University of Bari. She obtained her Ph.D. defending a dissertation on Lexical Semantics of Dialogue Acts, under the supervision of dr. Carlo Strapparava. Her work was financed in part by HUMAINE, the European Human-Machine Interaction Network on Emotion. Her interests are in the area of intelligent interfaces, natural language processing, affective computing, cognitive science, user models and collaborative software development.

Carlo Strapparava is a senior researcher at FBK-irst (Fondazione Bruno Kessler - Istituto per la ricerca scientifica e Tecnologica) in the Human Language Technologies Unit. His research activity covers AI, natural language processing, intelligent interfaces, cognitive science, knowledge-based systems, user models, adaptive hypermedia, lexical knowledge bases, word sense disambiguation, and computational humor. In these fields, he published more than 150 scientific reviewed publications.