Automatic Detection of Dialog Acts Based on Multi-level Information †*

Sophie Rosset and Lori Lamel
Spoken Language Processing Group, LIMSI-CNRS,
B.P. 133, 91403 Orsay cedex, France
{rosset, lamel}@limsi.fr
http://www.limsi.fr/tlp

ABSTRACT

Recently there has been growing interest in using dialog acts to characterize human-human and human-machine dialogs. This paper reports on our experience in the annotation and the automatic detection of dialog acts in human-human spoken dialog corpora. Our work is based on two hypotheses: first, word position is more important than the exact word in identifying the dialog act; and second, there is a strong grammar constraining the sequence of dialog acts. A memory based learning approach has been used to detect dialog acts. In a first set of experiments the number of utterances per turn is known, and in a second set, the number of utterances is hypothesized using a language model for utterance boundary detection. In order to verify our first hypothesis, the model trained on a French corpus was tested on a corpus for a similar task in English and for a second French corpus from a different domain. A correct dialog act detection rate of about 84% is obtained for the same domain and language condition and about 75% for the cross-language and cross-domain conditions.

1. INTRODUCTION

In order to capture the richness of human-human call center dialogs, it is interesting to explore and correlate dialog features at multiple levels: lexical, semantic and functional. We are also interested in automatically modeling the discourse structure in order to develop more sophisticated spoken dialog systems.

A useful analysis involves the identification of *dialog acts*. Dialog acts are functional abstractions over variations of utterance form and content. Some examples of dialog acts are *Assert, Information-Request, Acknowledgment*, meant to capture things speakers are attempting to do with speech. Many taxonomies of dialog acts have been proposed ([13]). One of the most complete and widely used is the DAMSL taxonomy. This tagging system has been used and adapted for a variety of projects, including the European and American project AMITIES (Automated Multilingual Interaction with Information and Services) project. In AMITIES a method for annotating dialogs at multiple levels has been developed based on the DAMSL scheme [5].

Some of the recent research on dialog has been based on the assumption that the dialog acts are good way to characterize dialog behaviors in both human-human and human-machine dialogs [1, 4, 7]. The work reported in [6] is driven by the observation that dialog acts are correlated with cue-phrases (or word substrings). This approach has the problem that words substring are often quite task and domain dependent. To overcome this problem [9] proposed using word n-grams. The approach proposed by [10] uses cue-phrases and a subset of dialog acts cues (word n-grams). Generally speaking, there is a many-to-many mapping between dialog acts and words. For example, the single word such *OK* could correspond to different dialog acts such as *backchannel, response, confirm*. On the other hand, the dialog act *assert* can be realized by many different word sequences (and utterances units) such *my birthdate is 02/23/65, 68 euros 50...* In order to reduce task dependence and to handle such multiple mappings, we are interested in finding a way to determine dialog acts without explicit use of lexical information, our hypothesis being that this information is not crucial. Thus, one of the main goals for this work was to examine what various kinds of information are useful for automatic dialog act (DA) tagging. In the remainder of this paper we describe our methodology for automatic dialog act tagging applied to corpora for two tasks and in two languages.

2. DIALOGIC ANNOTATION

A dialog can be divided into units called turns, in which a single speaker has temporary control of the dialog and speaks for some period of time. Within a turn, the speaker may produce several utterances units where the definition of an utterance unit is based on an analysis of the speaker's intention (the dialog acts). Once a turn is segmented into units, annotation involves making choices along several dimensions, each one describing a different orthogonal aspect of the utterance unit. The utterance tags summarize the intentions of the speaker and the content of the utterance unit. The taxonomy adopted in the AMITIES project follows the general DAMSL categories where the dialogic tags are classified into five broad categories:

- **Communicative Status:** refers to the features of the communication
- **Information Level:** characterizes the semantic content of the utterance unit
- **Forward Looking Function:** refers to how the current utterance unit constrains future beliefs and actions of the participants, and affects the discourse
- **Backward Looking Function:** refers to how the current utterance unit relates to the previous one
- **Conventional:** refers to utterance units which initiate or close the dialog

The dialog acts for **Communicative Status** are *Self-talk, Third-party-talk, Abandon, Interruption and Change of mind*. The **Information-Level** includes *Task, Task-management (System capabilities, Order of tasks, Completion and Summary),*

* † This work was partially funded by the IST-AMITIÉS project

Corpus	#dialogs	#turns	#words	#distinct	#utts
GE_fr	134	4273	37.8k	1473	5623
CAP_fr	24	1034	8.8k	1109	1359
GE_eng	21	2219	11.0k	750	2649

Table 1: Characteristics of the corpora.

Communication-management, and Out-of-topic. The role of the **Forward-Looking Function** is to anticipate the future in some way with dialog acts corresponding to *asking a question*, *making a statement*, *committing to an action*, or *telling the other person to do something*. **Backward-looking functions** are primarily responses *(agreement, answer and understanding)*. to a previous turn, often a response to a question. If some level of agreement or disagreement with the previous speaker's question (or some degree of accepting or rejecting the previous speaker's proposal) is signaled, then an *Agreement* tag is selected. Most *acceptances* and *rejections* are also answers. The *Understanding* tag is used to denote that some level of understanding or misunderstanding is signaled by the speaker. Because the dialogic tags cover several aspects of conversations, multiple labels are usually associated with a particular utterance unit. Every utterance unit may be categorized according to its information level and to its immediate function, which means that an utterance unit can be tagged with labels from all the categories. For instance, the utterance unit *A for Alpha* is labeled with the Forward-looking function *Explicit-Confirm-request* and the Backward-looking function *Non-understanding*.

3. CORPUS AND METHODOLOGY

Three corpora were used in this study (see Table 1). The first one (GE_fr) consists of agent-client dialogs in French recorded at a bank call center service. The dialogs cover a range of investment related topics such as information requests (credit limit, account balance), orders (change the credit limit) and account management (open, close, modify personal details). The application domain is structured into 6 major topics, hierarchically organized into 45 sub-topics. The first corpus was divided into 3 sets for training, development and testing purposes containing 94 (2623 turns, 3912 utterance units); 21 dialogs (675 turns, 869 utterance units) and 19 dialogs (675 turns, 842 utterance units) respectively. The second corpus (CAP_fr) consists of agent-client recordings in French from a Web-based Stock Exchange Customer Service center. While many of the calls concern problems in using the Web to carry out transactions (general information, complicated requests, transactions, confirmations, connection failures), some of the callers simply seem to prefer interacting with a human agent. The dialogs cover a range of investment related topics such as information requests (services, commission fees, stock quotations), orders (buy, sell, status), account management (open, close, transfer, credit, debit) and Web questions/problems. The third corpus (GE_eng) consists of agent-client dialogs in English recorded at a bank call center service. The dialogs cover essentially the same investment related topics as the GE_fr corpus. In addition to the AMITIES dialogic annotations, the training corpora were tagged with named entities (expressions for people, places, organizations, ...) and task entities. Task entities are named entities which describe task or domain specific knowledge such as *account number, account amount* etc. In this study, 2 of the 5 broad classes used in AMITIES have been further subdivided so as to allow multiple tags to be specified for each utterance unit: the Forward-looking function class is split into 2 sub-classes (Statement and Influence on Listener), and the Backward-looking function class was divided into 3sub-classes

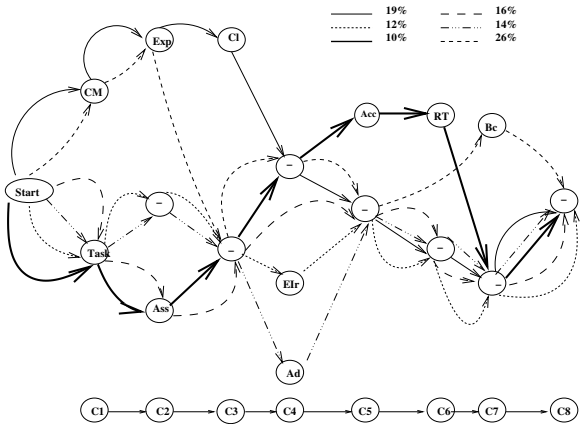

Figure 1: Most frequent successions of DAs (200+ turns).

(Agreement, Answer and Understanding). The new 44 dialog act tag taxonomy is:

- **Class1** *Information Level*: Communication-mgt, Out-of-topic, Task, Task-management-Completion, Task-management-Order, Task-management-Summary, Task-management-System-Capabilities
- **Class2** *Forward Looking Function - Statement*: Assert, Commit, Explanation, Expression, ReExplanation, Reassert
- **Class3** *Conventional*: Closing, Opening
- **Class4** *Forward Looking Function - Influence on Listener*: Action-directive, Explicit-Confirm-request, Explicit-Info-request, Implicit-Confirm-request, Implicit-Info-request, Offer, Open-Option, Re-Action-directive, Re-Confirm-request, Re-Info-request, Re-Offer
- **Class5** *Backward Looking Function - Agreement*: Accept, Accept-part, Maybe, Reject, Reject-part
- **Class6** *Backward Looking Function - Answer*: Response-To
- **Class7** *Backward Looking Function - Understanding*: Backchannel, Completion, Correction, Non-understanding, Repeat-rephrase
- **Class8** *Communicative Status*: AbandStyle, AbandTrans, AbandChangeMind, AbandlossIdeas, Interrupted, Self-talk

Since each utterance unit could potentially receive one tag for each of the 8 classes, the tags are represented by a vector with one item per class. If none of the class' tags is relevant it is represented by NA (not applicable). For example, the following utterance units have the corresponding dialog context-dependent tags:

Client: *four years*
DAs: *Task Assert NA NA NA Response-To NA NA*
Client: *[number]*
DAs: *Task NA NA Explicit-Confirm-request NA NA Repeat-rephrase NA*.

Only 197 different combinations of dialogs acts are observed in the 3912 training utterance units. There is a strong predictive factor in the succession class tags in the utterance unit. This is illustrated in Figure 1 which represents the 6 most frequent dialog act sequences (accounting for 51% of the training utterance units) found in the training data as a grammar. For example, if the Class1 tag is Task (52%), then the Class2 tag is either NA (26%) or Assert (26%), and Class3 is NA. Figure 2 shows an example of how the training data is represented. For each utterance unit of each speaker turn, an entry specifies the tag for each of the 8 dialog act classes. A Memory Based Learning methodology

Transcript: GE Capital Bank [name] *introduction
Utterance unit 1, Class2 Dialog Act
Agent 2 GE Capital Communication-mgt Assert
Utterance unit 2, Class2 Dialog Act
Agent 2 GE Capital Bank [name] Communication-mgt Assert
Opening NA NA NA NA NA +Communication-mgt Expression

Figure 2: Example dialog act annotations used for training. The example transcription is after named entity (*[name]*) and task entity (**introduction*) mapping. The tags following the + correspond to the second utterance unit.

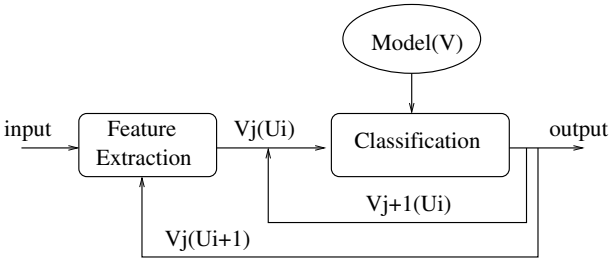

Figure 3: Similarity based reasoning using previous hypothesis

was adopted since such methods work well with small amounts of data and have been shown to be well adapted for natural language processing [2, 3]). We employed the IB1-IG implementation of Machine Based Learning from TiMBL software package ([12]) with the Manhattan distance, one of the most basic metrics which works well with symbolic features. In this metric, the distance between two patterns is simply the sum of the differences between the features. The feature weights used by the k-nearest neighbor (k-NN) algorithm are a gain ratio, (a normalized version of the Information Gain measure) computed from the training data. Our goal is to automatically detect the dialog acts for each speaker turn. Figure 3 schematically represents the dialog act classification method. According to our first hypothesis that word position is more important than the exact words, only the first words of each turn are used as lexical features. The identity of the speaker (Agent or Client) is also used. In order to take into account the dialog act grammar and the utterance unit context (the previous tags in utterance unit), the number of utterance units in each turn is used. The features include all previously proposed tags (for the completed tag classes along the utterance unit and all the previous hypotheses done on the previous utterance unit). At the first step, a speaker turn is input to the system which extracts the defined features (speaker identity, number of utterances and first 2 words) and puts them into a vector ($V_1(U_1)$=[SpkrId, #Utt., w_1, w_2]). The classification of the vector (e.g. assigning a dialog act to it) is done comparing the vector to all the examples in the training database. The result of this first classification is considered as an element of the vector used to classify the next dialog act ($V_i + 1(U_1)$. After the utterance has been classified for all 8 dialog acts classes, if there are more than one utterance unit in the turn, the next two words of the turn are added to the vector containing the hypotheses for the previous utterance ($V_i(U_i + 1)$. Figure 4 represents the method used to model the data, which consists of computing the weight of each feature for each type of vector.

4. EXPERIMENTS AND RESULTS

Detection of DAs: known number of utterance units

The first experiments were carried out using the GE_fr corpus, with a model trained on the designated training portion. The

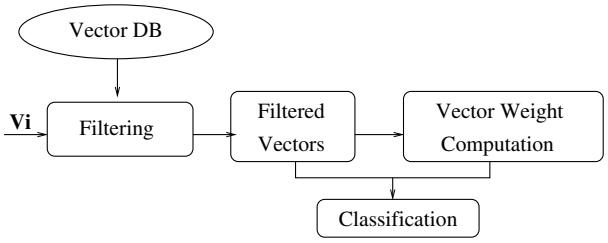

Figure 4: Estimation of Model(V)

Data	#dial	#utt.	#turn	%correct	condition
GE_fr dev	21	869	675	86.0	4words
				86.0	4+2words
				85.8	2+2words
GE_fr Test	19	842	675	83.3	4words
				83.5	4+2words
				83.4	2+2words

Table 2: Percent correct DA detection on GE_fr dev and test data with the GE_fr model for different experimental conditions.

number of words used in the input vector was varied: (1) using the first 4 words in the turn as the first utterance unit; (2) using 4 words in the first utterance unit and 2 more words for each subsequent utterance unit; (3) using 2 words in the first utterance unit and 2 more words for each subsequent one. The results are given in Table 2. An examination of the χ^2 measures indicated that most of the time words are really not relevant features. For example, the χ^2 value is 885.8 with 63 different features for the first word and 345.4 with 7 different features values for the first DA. This observation supports our hypothesis on the role of words.

To test the hypothesis further, the models trained on GE_fr were applied to the CAP_fr corpus (a change of task) and to the GE_eng corpus (a change of language). The results given in Table 3 are somewhat better for the 2+2words model, particularly for the English data. Considering that only 40% of first utterance units in the GE_eng data contain a named entity or a task entity, the 75% correct detection rate adds support to our initial hypothesis. However looking more closely at the 7 most frequent dialog acts (see Table 4) suggests that some of the dialog acts are more language and task dependent than others.

Detection of DAs: estimated number of utterance units

The results presented above assumed that the number of utterance units was known a priori. In these experiments an automatic method is used to hypothesize the number of utterances in the speaker turn. The method uses a 4-gram language model trained on the normalized transcripts (with named and task entities) to model the distribution of words and utterances boundaries (explicitly represented in the training data). The language model was used to predict the most probable locations of utterances boundaries, thereby providing an estimate of the number of utterance units. Table 5 shows segment detection results on the GE_fr data.

Data	# dial	#utt.	#turn	%correct	condition
CAP_fr	24	1208	1034	74.7	4words
				74.8	4+2words
				75.0	2+2words
GE_eng	21	1335	1109	70.1	4words
				70.1	4+2words
				75.5	2+2words

Table 3: DA detection rates using the GE_fr model on CAP_fr and GE_eng test data.

DA	GE_fr test	CAP_fr	GE_eng
Response-To	52.0% (125)	33.0% (184)	55.7% (458)
Backchannel	75.0% (142)	72.0% (162)	89.2% (148)
Accept	51.7% (143)	26.0% (131)	30.3% (108)
Assert	66.0% (233)	56.3% (320)	50.5% (540)
Expression	89.0% (343)	69.3% (408)	56.2% (137)
Comm-mgt	86.8% (395)	70.7% (479)	59.2% (444)
Task	85.4% (397)	81.4% (529)	78.8% (864)

Table 4: Detection results on the 7 most frequent DAs.

data	#utt	#turn	%turn INS	%turn DEL	%utt INS	%utt DEL
dev	868	674	5.4	7.0	4.3	5.9
test	841	674	4.0	7.5	3.3	6.6%

Table 5: Detection of the number of utterance units in GE_fr.

Roughly 5-7% of the speaker turns have an utterance insertion or deletion, with an overall utterance boundary insertion rate of 4% and a deletion rate of about 6%. The hypothesized number of utterance segments was used for dialog act detection in place of the known number of utterance units. Table 6 gives results on the GE_fr dev and test data in terms of DA insertions, deletions, and substitutions.

5. CONCLUSION AND PERSPECTIVES

This paper has reported on recent work with automatic dialog act tagging for different corpora, as well as automatic detection of the number of utterance units. Starting with the AMITIES multilevel dialog annotations based on DAMSL, a set of 8 dialog act classes were defined. Each utterance unit is represented by a vector with values (which can be empty) for each of the 8 DA classes. A Memory Based Learning approach was used to compare the feature vectors of the test data to those in the training data. The features include the speaker, the number of utterance units in the turn, the previous (hypothesized) dialog acts and 2 words per utterance unit. When the number of utterance units is known, the DA detection rate is about 83% in the same task/language condition. Using the model under cross-domain and cross-language conditions resulted in a DA detection rate of about 75%, lending support to our underlying hypothesis that the position of a word is more important that the exact word as a predictor of the dialog act. However, an analysis of the most frequent errors shows that some dialog acts are more task or language dependent than the global results suggest.

In order to automatically detect the dialog acts and model the dialog structure, the utterance unit boundaries need to be automatically located (however, for this work only the number of utterances in each turn is used). A 4-gram language model was used to predict the most probable locations of utterance boundaries (and thereby the number of utterance units) in each turn. For about 88% of speaker turns, the utterance boundaries were correctly detected. The loss in DA accuracy arising from use of the automatically detected utterance unit boundaries as features instead of the manually located ones is primarily due to insertion and deletion errors, with the substitution rate remaining similar to the known utterance unit boundary condition.

The data show that there is a strong grammar between the dialog act classes in a single utterance unit and a strong grammar between 2 or more utterance units. It also appears that the data normalization (task and named entities) reduces the language dependency of this approach. It is likely that other sources of information such as the dialog history could also be useful to predict dialog acts. Our belief is that modeling the richness of human dialogs may lead to the development of more sophisticated spoken dialog systems.

data	# DAs	%INS	%DEL	%SUBS
dev	6944	4.3	5.9	10.6
test	6728	3.3	6.6	13.5

Table 6: Automatic detection of DAs using hypothesis on number of utterance units

6. ACKNOWLEDGMENTS

We would like to offer special thanks to Isabelle Wilhelm for her participation in the specification of the annotation conventions and the dialog annotations without which this work would not have been possible.

REFERENCES

[1] R. Cattoni, M. Danieli, A. Panizza, V. Sandrini, C. Soria. 2001. "Building a corpus of annotated dialogues: the ADAM experience," *Corpus-Linguistics-2001*.

[2] A. van den Bosch, E. Krahmer, M. Swerts. 2001. "Detecting problematic turns in human-machine interactions: Rule-induction versus memory-based learning approaches," *ACL'00*,pp. 499-506.

[3] W. Daelemans, A. van den Bosch, J. Zavrel. 1999. "Forgetting exceptions is harmful in language learning," Machine Learning, **34**:11-43.

[4] B. Di Eugenio, P. W. Jordan, J. D. Moore, R. H. Thomason. 1998. "An empirical investigation of collaborative dialogues," *ACL-COLING98*.

[5] H. Hardy, K. Baker, H. Bonneau-Maynard, L. Devillers, S. Rosset, T. Strzalkowski. 2002. "Semantic and Dialogic annotation for Automated Multilingual Customer Service," ISLE workshop.

[6] J. Hirschberg, D.J. Litman. 1993. "Empirical Studies on the Disambiguation of Cue Phrases," *Computational Linguistics,* **19**(3):501-530.

[7] A. Isard, J.C. Carletta. 1995. "Replicability of transaction and action coding in the map task corpus," *AAAI Spring Symposium: Empirical Methods in Discourse Interpretation and Generation*, pp. 60-67.

[8] R. Prasad, M. Walker. 2002. "Training a Dialogue Act Tagger For Human-Human and Human-Computer Travel Dialogues," *Third SIGdial Workshop on Discourse and Dialogue*, pp. 162-173.

[9] N. Reithinger, M. Klesen. 1997. "Dialogue act classification using language models," *Eurospeech'97*,pp. 2235-2238.

[10] K. Samuel, S. Carberry, K. Vijay-Shanker. 1998. "Dialogue act tagging with transformation-based learning," *COLING-ACL*, pp. 1150-1156.

[11] E. Shriberg, P. Taylor, R. Bates, A. Stolcke, K. Ries, D. Jurafsky, N. Coccaro, R. Martin, M. Meteer, C. Van Ess-Dykema. 2000. "Can prosody aid the automatic classification of dialog acts in conversational speech," *Language and Speech*.

[12] W. Daelemans, J. Zavrel, K. van der Sloot, A. van den Bosch. 2003. *TiMBL: Tilburg Memory Based Learner, v5.0, Reference Guide, ILK Technical Report ILK-03-10*, (http://ilk.kub.nl/software.html#timbl)

[13] D. Traum. 2000. "20 Questions on Dialog Act taxonomies",*Journal of Semantics*, **17**(1):7-30.