Multi-level Information and Automatic dialog Acts Detection in human-human Spoken Dialogs

Sophie Rosset and Delphine Tribout

Spoken Language Processing Group, LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France
{rosset,tribout}@limsi.fr
http://www.limsi.fr/tlp

1. Introduction

Recently there has been growing interest in using dialog acts to characterize human-human and human-machine dialogues. In order to capture the richness of human-human call center dialogues, it is interesting to explore and correlate dialog features at multiple levels: lexical, semantic and functional. We are also interested in automatically modeling discourse structure in order to develop more sophisticated spoken dialogue systems.

A useful analysis involves the identification of dialog acts. A dialog can be divided into units called turns, in which a single speaker has temporary control of the dialog. Within a turn, the speaker may produce several utterance units. An utterance unit is a continuous segment of speech covering one intention. Once a turn is segmented into units, annotation involves making choices along several dimensions, representing different orthogonal aspects of the utterance unit. For instance, one dimension characterizes the effect the utterance has on the other speaker, such as a request for information or making a statement. Another dimension shows that a speaker has understood what has been said to him or her. A dialog act represents a value along one of these dimensions like understanding. Each utterance unit receives several dialog acts often called tags.

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 ([18]). 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

2. Corpus and Methodology

2.1. Corpus

The main corpus (GE-fr) used in this study consists of 134 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. These dialogs were orthographically transcribed with Transcriber, a tool for segmenting, labeling and transcribing speech [3]. This corpus was divided into 2 sets for training (containing 94 dialogs, 2923 turns, 3912 utterance units) and testing purposes (containing 40 dialogs, 1350 turns, 1711 utterance units). For the second part of our experiments, we used two other corpora: the first one, CAP_fr (24 dialogs, 1025 turns, 1203 utterance units), 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 European and American project AMITIES (Automated Multilingual Interaction with Information and Services) project.

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 dialogues [4, 7, 10]. The work reported in [9] is driven by the observation that dialog acts are correlated with cue-phrases (or word substrings). This approach has the problem that word substrings are often quite task and domain dependent. To overcome this problem [13] proposed using word n-grams. The approach proposed by [14] uses cue-phrases and a subset of dialog act cues (word n-grams). Generally speaking, there is no usable mapping between tags and words. For instance, the single word such OK could correspond to different dialog acts such as backchannel, response, confirm. On the other hand, a dialog act such as assert can be realized by many different word sequences like I am 34, 8 euros 50. In light of the above, 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 critical. 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.

The remainder of this paper is organized as follow: Section 2 introduces the three corpora used, our methodology and the basic system [16]. Section 3 describes the different experiments and gives results, followed by conclusions in Section 4.
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 dialog covers 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 second one, GE.eng (31 dialogs, 1147 turns, 1357 utterance units), 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.

2.2. Dialogic Annotation

Once a turn is segmented into units, these have to be annotated in dialog acts. The taxonomy adopted in the AMITIES project [8] follows the general DAMSL categories [2]. In AMITIES a method for annotating dialogs at multiple levels was developed based on the DAMSL scheme. In this study, the dialogic tags are classified into eight dimensions to allow multiple tags to be specified for each utterance unit (if no tag is relevant it is represented by NA (not applicable)).

- **Class 1 Information Level**: characterizes the semantic content of the utterance unit. The different tags are Communication-mgt, Out-of-topic, Task, Task-management-Completion, Task-management-Order, Task-management-Summary, Task-management-System-Capabilities.
- **Class 2 Statement**: makes a claim about the world, and tries to change the beliefs of the listener. The different tags are Assert, Commit, Explanation, Expression, ReExplanation, Reassert.
- **Class 3 Conventional**: refers to utterance units which initiate or close the dialog. The different tags are Closing, Opening.
- **Class 4 Influence on Listener**: In this group of tags, the speaker is asking the listener a question, directing him or her to do something, or suggesting some course of action the listener may take. The different tags are 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.
- **Class 5 Agreement**: indicates whether the speaker accepts a proposal, offer or request, or confirms the truth of a statement or confirmation-request. The different tags are Accept, Accept-part, Maybe, Reject, Reject-part.
- **Class 6 Answer**: is a response to an Information-request or Confirmation-request. An answer by definition will always be an assertion, as it provides information or confirms a previous supposition, and it makes a claim about the world. Therefore only one tag is used: True.
- **Class 7 Understanding**: reveals whether and in what way the speaker heard and understood what the other speaker was saying. The different tags are Backchannel, Completion, Correction, Non-understanding, Repeat-rephrase.
- **Class 8 Communicative Status**: refers to the features of the communication. The different tags are AbandStyle, AbandTrans, AbandChangeMind, Aband-lossIdeas, Interrupted, Self-talk.

Even if the number of possible tag combinations is huge (1,016,064), only 197 are observed in the 3912 training utterance units. Six of them represent 51% of the corpus. For example, if the Class1 tag is Task (52%), then the Class2 tag is either NA (26%) or Assert (26%), and Class3 is NA. There is a strong predictive factor in the succession class tags in the utterance unit.

2.3. Methodology: basic system

This work is based on three hypotheses:

1. The Dialog Act succession is strongly constrained as previously shown.
2. The initial words are more important than the remaining words in identifying the dialog act for example I’d like... can you give me...
3. The information is encoded in specific entities:
   - Named Entities which are expressions for people, places, organizations
   - Task Entities which are named entities which describe task or domain specific knowledge such as account number, account amount
   - Linguistic Entities which give structure to the utterances, for example I’d like to...

Our goal is to automatically detect the dialog acts for each speaker turn. All the data have been automatically tagged with specific entities. This tagging is done in two steps: the first one is language dependent but task independent and consists of automatic tagging of named entities. The second one is language independent but task dependent, consisting of task entity detection. These taggers use rewrite rules which work like local grammars and with specific dictionaries. They replace the specific entities by tag words expressing their types. First, the turn is tagged. Each speaker turn is input independently to the system. The N first words of each turn are used as lexical features. The identity of the speaker (Agent or Client) and the number of utterance units in the turn are used as additional information. All these features are put in a vector and the dialog act for the first dimension is predicted using memory based learning (MBL), specifically the Timbl implementation ([17]), since it works well with small amounts of data and it has been shown to be well adapted for natural language processing ([5, 6]). We use the Manhattan distance, where the distance between two patterns is simply the sum of the differences between the features. MBL works by finding the vector in the training database closest to the test one. The result of this first prediction is considered as an element of the vector used to predict the next dialog act ($DA_i(UU_{j})$=$[SpkrId, \#Utt., w_1,...,w_{N},DA_{i-1}]$). After the utterance has been classified for all 8 dialog act dimensions, if there is more than one utterance unit in the turn, the N next words of the turn are added to the vector containing the hypotheses for the previous utterance unit ($DA_i(UU_j)$=$[SpkrId, \#Utt., DAs(UU_{j-1}) + DAs(UU_{j-1} + w_{uu_{j}} w_{uu_{j}})]$). For example, the turn Agent: donnez-moi votre numéro de compte (give me your account number) having the following dialog acts tags: DAs: information-level=Task; influence-on-listener=Action-directive is represented by the following vector:

Agent 1 donnez-moi votre numéro Task NA NA Action-directive NA NA NA
And the turn Client: alors [numerique] [numerique] [numerique] [numerique] [numerique] [numerique] (then [number] [number] [number] [number] [number] [number] having the following...
3. Experiments and Results

3.1. Utterance unit boundary information

The first experiments were carried out using the GE\textsubscript{fr} corpus, with a model trained on the designated training portion. We used, instead of the N first words of the turn, the N first words of each utterance unit to be tagged. The number of words in the input vector was varied: (1) using the first 4 words in 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. According to our hypothesis that information is encoded in specific entities, specific entities were used instead of exact words. In addition, we used different models for the Agent and the Client. To further test our hypothesis on the role of words, the models trained on GE\textsubscript{fr} were applied to the CAP\textsubscript{fr} corpus (a change of task) and to the GE\textsubscript{eng} corpus (a change of language) with the same conditions. The results given in Table 1 are somewhat better for the 2+2words model, particularly for the English data. These results are slightly better than those reported using no utterance unit information and a general model.

3.2. Use of the historical information

These experiments using historical information are based on two hypotheses: the first one is that there are relations between the different utterance units in one turn and that these relations are organized; the second one is that a dialog being a succession of turns, the dialog acts of a turn have an incidence on the dialog acts of the next turn. The first hypothesis is based on some observations of the results of our first model. These results indicate that the dialog history, or more precisely, the position of the utterance unit has a different role and weight. Table 2 shows the different dialog act detection error rates for each position of utterance unit. The tags of the previous utterance unit added to the vector used to annotate the current utterance unit are seen to only be helpful for tagging the second utterance unit. (We also tried always using just the previous UU tags as the history, but the results were also worse for UU 3 and UU 4.)

Table 3 (exp. 1) gives results for the 4+2words condition, where the dialog history is not used to annotate the third utterance unit. The DA error rate is seen to be lower than in Table 2. In order to capture a larger dialog history, we decided to add the dialogic information of the previous turn. Based on the previous results, only the information of the last utterance unit of the previous turn has been used. The results for the first two utterance units are seen to improve using this history (see Table 3, exp. 2). The third experiment is a mix of the two previous ones. For the first two utterance units, the dialogic information of the last utterance unit of the previous turn was used. The third utterance unit of the current turn is considered as a first utterance unit and no previous history information is added to the vector. The results are better for all the utterance units (Table 3, exp. 3). The results for utterance unit 4 are not very reliable because there are very few turns with 4 utterance units in the data (17 in training; 6 in test). We can observe that the third experiment give the best results for all the utterances and the global turn. In looking closely at the data we observed that when a turn contains

<table>
<thead>
<tr>
<th>Data</th>
<th># Dial</th>
<th># Turn</th>
<th># Utt.</th>
<th>% Error 2+2w</th>
<th>% Error 4w</th>
<th>% Error 4+2w</th>
<th>Expt.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GE\textsubscript{fr}</td>
<td>134</td>
<td>1350</td>
<td>1711</td>
<td>15.0</td>
<td>14.4</td>
<td>14.6</td>
<td>4words</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4+2words</td>
<td></td>
<td></td>
<td>2+2words</td>
</tr>
<tr>
<td>Cap\textsubscript{fr}</td>
<td>24</td>
<td>1025</td>
<td>1203</td>
<td>22.8</td>
<td>22.3</td>
<td>22.8</td>
<td>4words</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4+2words</td>
<td></td>
<td></td>
<td>2+2words</td>
</tr>
<tr>
<td>GE\textsubscript{eng}</td>
<td>51</td>
<td>1147</td>
<td>1357</td>
<td>30.4</td>
<td>30.2</td>
<td>25.3</td>
<td>4words</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4+2words</td>
<td></td>
<td></td>
<td>4+2words</td>
</tr>
</tbody>
</table>

Table 1: Error rate for DA detection on GE\textsubscript{fr}, GE\textsubscript{eng} and Cap\textsubscript{fr} test data with the GE\textsubscript{fr} model for different experimental conditions.

<table>
<thead>
<tr>
<th># Utt.</th>
<th>%Error 2+2w</th>
<th>%Error 4w</th>
<th>%Error 4+2w</th>
</tr>
</thead>
<tbody>
<tr>
<td>UU 1</td>
<td>15.0</td>
<td>14.7</td>
<td>14.7</td>
</tr>
<tr>
<td>UU 2</td>
<td>12.6</td>
<td>16.0</td>
<td>12.4</td>
</tr>
<tr>
<td>UU 3</td>
<td>18.0</td>
<td>18.7</td>
<td>19.8</td>
</tr>
<tr>
<td>UU 4</td>
<td>20.8</td>
<td>18.7</td>
<td>20.8</td>
</tr>
<tr>
<td>Turn</td>
<td>14.6</td>
<td>15.0</td>
<td>14.4</td>
</tr>
</tbody>
</table>

Table 2: Dialog Act Error Rate and Utterance Unit in GE\textsubscript{fr} test data using dialog history.

<table>
<thead>
<tr>
<th>Exp1</th>
<th>Exp2</th>
<th>Exp3</th>
</tr>
</thead>
<tbody>
<tr>
<td>UU 1</td>
<td>14.7</td>
<td>11.9</td>
</tr>
<tr>
<td>UU 2</td>
<td>12.4</td>
<td>13.6</td>
</tr>
<tr>
<td>UU 3</td>
<td>14.7</td>
<td>20.9</td>
</tr>
<tr>
<td>UU 4</td>
<td>20.8</td>
<td>18.7</td>
</tr>
<tr>
<td>Turn</td>
<td>14.3</td>
<td>12.5</td>
</tr>
<tr>
<td></td>
<td>12.3</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Error rate of dialog act detection with different uses of historic information (4+2words condition).
The same experiments were carried out on the Cap\textsubscript{fr} and GE\textsubscript{eng} test data, using the GE\textsubscript{fr} model, using that 2+2 words condition which gave the best results with the baseline model. The results are summarized in Table 4. For the cross-domain and the cross-language conditions, the results are still better with the historical model than with the basic model.

### 4. Conclusion and Perspectives

This paper has reported on recent work with automatic dialog act tagging for different corpora. Starting with the AMITIES multilevel dialog act annotations based on DAMSL, a set of 8 dialog act classes were defined. A Memory Based Learning approach was used to compare the feature vectors of the test data to those in the training data. The basic system uses as features the speaker, the number of utterance units in the turn, the previous (hypothesized) dialog acts and N words (or Specific Entities) per utterance unit. It also appears that the data normalization with specific entities reduces the language and task dependency of this approach. With the basic model (word-based features), the dialog act detection error rate was about 16%, and about 15% for the tag-word-based model. The data show that there are strong constraints between the dialog act dimensions in a single utterance unit and between successive turns. After an analysis of the results of the basic system, some dialogic information were added to the vector. These dialogical information are based on the analysis of the structure of the dialogs. With our best model, the dialog act detection error rate is about 12.3% in the same task/language condition. Using the same model under cross-domain or cross-language conditions resulted in a dialog act detection error rate of about 20%. These results support our underlying hypotheses that most of the information is encoded in specific entities and that the dialogic structure is a crucial information to predict dialog acts. The experiments and results reported in this paper assumed that the number of utterance units and the localization of the boundary were known a priori. In order to automatically detect the dialog acts and model the dialog structure, the utterance unit boundaries need to be automatically located. Previous experiments reported in [16] show that using a simple 4-gram language model can be useful to predict the number of utterance but that the correct localization of the boundary was less good. It is likely that other sources of information such as acoustic or prosodic informations could also be useful to predict the boundary localization and the dialog acts. Some studies [1, 11] have reported these information to be useful.

### 5. References


