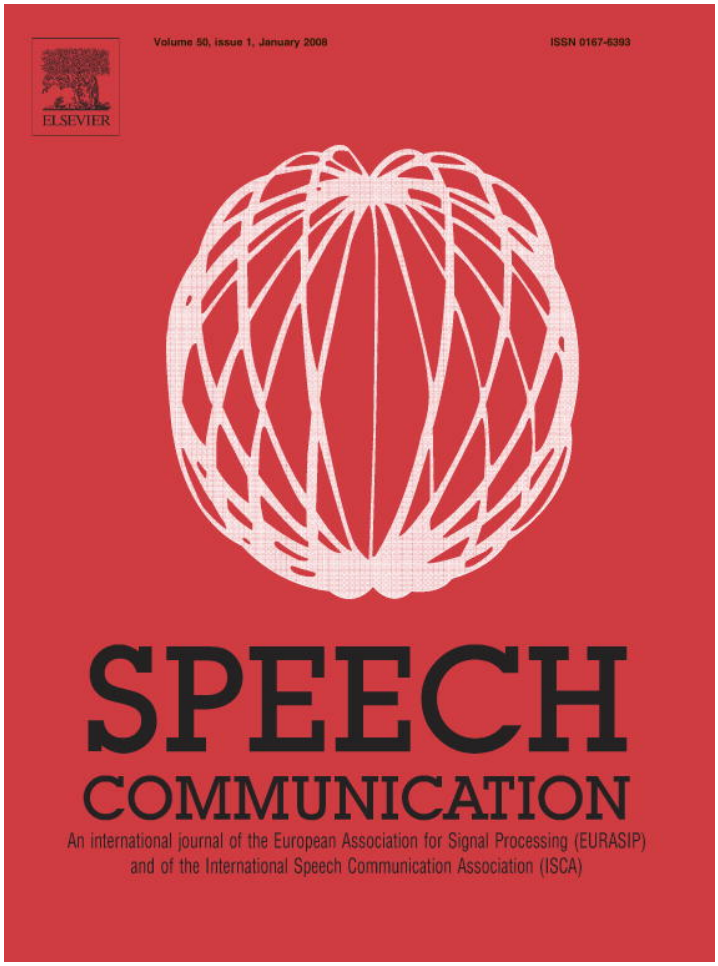

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 50 (2008) 1–13
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Multi-level information and automatic dialog act detection in human–human spoken dialogs

S. Rosset *, D. Tribout, L. Lamel

Spoken Language Processing Group, LIMSI-CNRS, F-91403 Orsay Cedex, BP 133, France

Received 14 August 2006; received in revised form 28 May 2007; accepted 28 May 2007

Abstract

This paper reports studies on annotating and automatically detecting dialog acts in human–human spoken dialogs. The work reposes on three hypotheses: first, the succession of dialog acts is strongly constrained; second, the initial word and semantic class of word are more important for identifying dialog acts than the complete exact word sequence of an utterance; third, most of the important information is encoded in specific entities. A memory based learning approach is used to detect dialog acts. For each utterance unit, eight dialog acts are systematically annotated. Experiments have been conducted using different levels of information, with and without the use of dialog history information. In order to assess the generality of the method, the specific entity tag based model trained on a French corpus was tested on an English corpus for a similar task and on a French corpus from a different domain. A correct dialog act detection rate of about 86% is obtained for the same domain/language condition and 77% for the cross-language or cross-domain conditions.

Keywords: Automatic dialog act detection; Human–human dialogs; Memory based learning

1. Introduction

Recently there has been growing interest in using dialog structure to characterize human–human and human–machine dialogs. One of the goals of such an analysis is to be able to automatically model discourse structure, with the hope of developing more sophisticated spoken dialog systems. In order to capture the complexity of human–human call center dialogs, it is interesting to explore and correlate dialog features at multiple levels: lexical, semantic and functional. Most dialog systems exploit the information present at the lexical and semantic levels. At the functional level the dialog can be described by a series of *dialog acts*.

Dialog acts attempt to capture things speakers are attempting to do with speech. Some examples of dialog acts are *Assert, Information-Request, Acknowledgment*. While many taxonomies of dialog acts have been proposed (Traum, 2000), one of the most complete and widely used is the DAMSL taxonomy (Allen and Core, 1997). This tagging system has been adapted for a variety of projects, including the joint European/American project AMITIES (Automated Multilingual Interaction with Information and Services) project (Hardy et al., 2003), the project in which this work was initiated.

Some of the recent research on dialog has been based on the assumption that the dialog acts are a good way to characterize dialog behaviors in both human–human and human–machine dialogs (Cattoni et al., 2001, 1998, 1995). Generally speaking, there is no unique mapping between dialog act tags and words. For instance, the single word "OK" could correspond to different dialog acts such as a *backchannel*, an *answer* to a question, or a *confirmation*. On the other hand, a dialog act such as *assertion* can be realized by many different word sequences: "I am 34" or "8 euros 50". In light of this lack of a direct

* Corresponding author.
E-mail addresses: rosset@limsi.fr (S. Rosset), tribout@limsi.fr (D. Tribout), lamel@limsi.fr (L. Lamel).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.05.007

correspondence between words and dialog acts, and in order to be as task and domain independent as possible, this work aims to find a way of determining dialog acts without the explicit use of lexical information, our hypothesis being that word (or multi-word expression) classes, for instance *balance_request*, are sufficient. 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 contrast to most reported work which annotate a single dialog act per utterance unit, in this work eight dialog acts are annotated for each utterance unit.

The remainder of this paper is organized as follows: Section 2 presents previous and related work. Sections 3 and 4 describe the three corpora used in this study and the dialogic annotation and classification scheme. Section 5 presents the methodology and Section 6 describes the different experiments carried out along with the results. Section 7 presents a further analysis of the results, followed by conclusions in Section 8.

2. Related work

Direct quantitative comparison of the results presented in this paper with other related work is not possible due to differences in the corpora and the annotation schemes used. Despite this, in this section we attempt to compare the different approaches at a conceptual level. This discussion highlights some of the alternative approaches proposed for dialog act tagging, specifying the corpora and annotations used, and discusses some differences with the approach adopted in this work. While most studies have been conducted using specific tasks (including this work), there has been growing interest in using corpora that are not linked to a specific human–machine or human–human interaction task.

Standard techniques from statistical language modeling have been applied to the dialog act tagging task. One of the most common approaches uses *n*-grams to model the probabilities of DA sequences. Nagata (1992) first proposed this approach and applied it on the ATR Conference Corpus Ehara et al. (1990). This corpus contains simulated dialogs between a secretary and a questioner at international conferences. The annotation scheme contains nine DAs and 2450 utterances. The nine DAs are *phatic, expressive, response, promise, request, inform, questionif, questionref, questionconf*. The model proposed by Nagata uses bigrams and trigrams conditioned by the preceding DAs to predict the upcoming DAs, and a tagging accuracy of about 40% was reported.

Other works have also relied on this approach (*n*-grams of DA) and proposed enhancements. For example, Reithinger et al. (1996) applied such an approach to the Verbmobil corpus. The Verbmobil task concerns meeting arrangements and trip planning. The corpus contains two-party scheduling dialogs and has been tagged with 43 DAs, grouped into 18 high level DAs. These 18 DAs are: *accept, bye, clarify, confirm, deliberate, digress, feedback, garbage, give_reason, greet, init, introduce, motivate, reject, request_comment, request_suggest, suggest* and *thank*. The complete annotation scheme is described in (Jekat et al., 1995). Reithinger et al. (1996) used a deleted interpolation to smooth the dialog act *n*-grams, reporting a tagging accuracy of 40% for the 18 high level DAs. Chu-Carroll (1998) incorporated knowledge of discourse structure in a corpus based study of airline reservation dialogues between two humans (SRI Transcripts, 1992). The corpus is comprised of 8 dialogs, 6 for training and 2 for evaluation. The dialog acts are: *Inform, Request-Referent, Answer-Referent, Request-If, Answer-If, Confirm, Clarify, Elaborate, Request-Explanation, Request-Repeat, Express-Surprise, Accept, Reject, Prompt, Greetings*. When different dialog acts could be applied to an utterance, the most specific was chosen, so that there was only one DA per utterance. Use of the Discourse Structure Information was shown to slightly improve system performance, and the tagging accuracy reported in this work is about 50%.

Stolcke et al. (2000) applied a somewhat more complicated hidden Markov model (HMM) method to the Switchboard corpus of conversational telephone speech. This corpus is the biggest corpus for which a DA study has been reported, containing 198k utterances (1155 dialogs) tagged with 42 DAs (1 tag/utterance). The annotation scheme, which is a simplification of the DAMSL tag set, is fully described in (Jurafsky et al.). The most frequent DAs are *Statement, Opinion, Yes-no-question, Declarative-question, Wh-question, Backchannel, Turn-exits, Abandonned-utterances, Yes-answers, No-answers, Agreement/Accepts, Reject and Maybe/Accept-part*. The method used by Stolcke et al. (2000) models both the sequencing of words within utterances and the sequencing of dialog acts over utterances. A tagging accuracy of 71% on the reference word transcripts of the Switchboard corpus is reported.

Other studies have investigated the use of cue phrases or word substrings for DA detection, for example Hirschberg and Litman (1993). This approach has the problem that word substrings are usually task and domain dependent. To overcome this problem, Reithinger and Klesen (1997) proposed using word *n*-grams and reported a tagging accuracy of 74.7% on the Verbmobil corpus. The approach proposed by Samuel et al. (1998) uses frequent word substrings. They used a modified version of the transformation-based learning (TBL Brill, 1995) over different utterance features such as utterance length, speaker turn and the dialog act tags of adjacent utterances. They tried different measures to select the cue phrases, and reported a tagging accuracy of 71.2% using entropy minimization with filtering and clustering on the Verbmobil corpus. In (Samuel et al., 1999), the same authors investigated a series of different measures for automatic selection of cue phrases, evaluating, among others measures, co-occurrence score, conditional probability, entropy, mutual information and deviation conditional probability (DCP). The best reported result uses the DCP metric (which measures how far a phrase deviates from an optimally predictive phrase),

which provided a tagging accuracy of about 71.5%. Webb et al. (2005) used a predictivity criterion not considered by Samuel et al. (1999) and reported a tagging accuracy of about 71.3% on the Switchboard corpus.

Dialog act classification has also been carried out on the Map Task corpus (Anderson et al., 1991). This corpus consists of conversations between two speakers, each having in their possession a different map of an imaginary territory. The task for the first speaker is to help the second one to draw a route only given on his/her map without access to the map of the second speaker. The corpus has been tagged with 12 different *types of utterance* (i.e. DAs) which result from a segmentation into turns. The 12 different DAs are *instruct, explain, align, check, query-yn, query-w, ack, clarify, reply – y, reply – n, reply – w, ready* and a thirteenth tag for uncodable segments has been added. Surendran and Levow (2006) applied support vector machine (SVM) and HMM approaches to detect dialog acts in this corpus. They reported 42.5% tagging accuracy using acoustic features, 59.1% using text features and 65.5% using both sets conjointly.

Ji and Bilmes (2005) proposed the use of dynamic Bayesian networks for tagging DAs. They worked with the ICSI meeting corpus (Janin et al., 2003), comprised of 75 meetings of a research group with an average of six speakers per meeting. The annotation scheme is close to the one used on the Switchboard corpus. Two methods were explored: switching *n*-gram models and factored language models. The factored language model method performed better, achieving a tagging accuracy of 66%.

The best results in the previous works were obtained with large amounts of training data which are quite expensive to produce. All of these studies also have in common the fact that they try to detect one DA per utterance. Our objectives in this work were to answer the following questions:

- Is there a simple way to do an automatic DA annotation using a small amount of training data and still get reasonable results?
- Is there an approach which could offer cross-domain or cross-language capabilities?
- Is it possible to detect all the different tags representing the different dimensions for each utterance?

With these objectives in mind, this work aims to find a way to determine dialog acts with a reduced use of lexical information, our hypothesis being that this information is not critical. Thus, this work examines what kinds of information are useful for automatic dialog act tagging, and explores the application of an approach known to work well when the amount of training data is small, the *Memory Based Learning* method.

3. Corpus

The data used in this study are from three sets of call-center dialogs recorded in the context of the AMITIES project (Hardy et al., 2003). The characteristics of the data are summarized in Table 1. The main corpus (GE_fr) contains 134 agent-client dialogs[1] 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 six major topics, hierarchically organized into 45 subtopics. The two other corpora were used to test the task and language portability of the method. The second corpus, CAP_fr, is comprised 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 in Leeds. The dialogs cover essentially the same investment related topics as the GE_fr corpus.

Table 1
Characteristics of the three corpora used in this study

	GE_fr	GE_eng	CAP_fr
# dialogs	134	31	24
# turns	4273	1147	1025
# turns/dialogs	32	37	43
# utterance units	5623	1357	1303
# utterance units/dialog	42	44	50
# utterance units/turn	1.3	1.2	1.2
# distinct words	1647	764	1123
# words	34336	6085	7741

The GE_fr data were divided into subsets for model development and for test. The GE_eng and CAP_fr data were used to test respectively the cross-language and cross-task capacity of the method.

The French corpora were orthographically transcribed with Transcriber, a tool for segmenting, labeling and transcribing speech (Barras et al., 2001). The English data were transcribed using a standard text editor. All three corpora were annotated with dialog acts using XDML Tool (eXtensible Dialogue Markup Language Tool) (Hardy et al., 2003).

4. Dialogic annotation

A dialog can be divided into segments called turns, in which a single speaker has temporary control of the dialog

[1] Only the portion of the AMITIES data with full multi-level annotations (Hardy et al., 2003) was used in this study. The full corpora contain 1067, 342 and 656 dialogs for GE_fr, CAP_fr and GE_eng, respectively. Different subsets of the corpora are annotated at multiple levels (topic, semantic, emotion).

and speaks for some period of time. Within a turn, the speaker may produce one or more 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 which cover a single intention, these are annotated with dialog acts. Annotation involves making choices along several dimensions, each one describing a different orthogonal aspect of the utterance unit. The dialog acts represent different aspects of an utterance. For instance, one dimension characterizes the effect an utterance has on the other speaker, such as a request for information or the making of 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 the dimensions, often referred to as a tag. The utterance tags summarize the intentions of the speaker and the content of the utterance unit. An example of the segmentation and annotation process is shown in Fig. 1. In this figure, the turn contains two utterance units, one introducing the service/agent and the other welcoming and implicitly passing control to the caller.

The taxonomy adopted in this work is derived from that adopted by the AMITIES project (Hardy et al., 2003) and follows the general DAMSL categories (Allen and Core (1997)) in which the dialogic tags are classified into five broad categories. In this study, two of the five broad classes have been further subdivided so as to allow multiple tags to be specified for each utterance unit: the Forward-looking function class was split into two subclasses (**Statement and Influence-on-Listener**), and the Backward-looking function class was divided into three subclasses (**Agreement, Answer and Understanding**). The resulting dialog act taxonomy has the following 8 classes and 44 tags:

- **Class 1 Information Level**: characterizes the semantic content of the utterance unit. The tags are *Communication-management, 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 influence the beliefs of the listener. The tags are *Assert, Commit, Explanation, Expression, Reassert, ReExplanation*.
- **Class 3 Conventional**: refers to utterance units which initiate or close the dialog. The possible 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 a 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 possible 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 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, AbandLossIdea, Interrupted, Self-talk.*

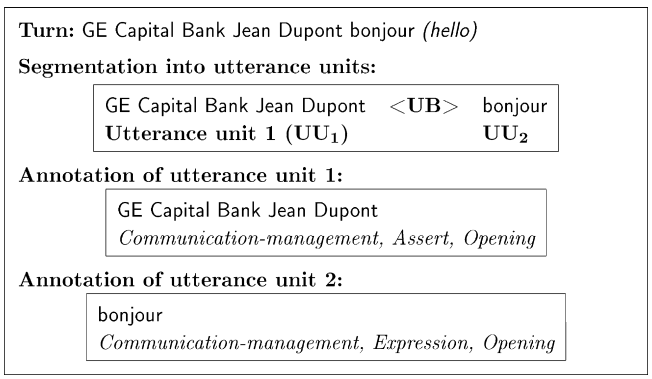

Fig. 1. Example of segmentation and annotation process. A Turn is first segmented into utterance units. Then these utterance units are annotated in dialog acts. The utterance unit boundary is denoted with the tag ⟨UB⟩.

Because the dialogic tags cover several aspects of the conversation, 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 potentially be tagged with labels from all of the categories. For instance, the utterance unit "A for Alpha" is labeled with the Influence-on-Listener tag *Explicit-Confirm-request* and the Understanding tag *Non-understanding*. If none of the tags for a class is relevant, it is labeled as NA (not applicable).

Although the number of possible tag combinations is huge (1,016,064), only 197 are observed in the 3912 training utterance units. Fig. 2 gives the six most frequent combinations of dialog acts found in the training data. These six combinations account for 51% of the training utterance units. For example, if the Class 1 tag is Task (54%), then the Class 2 tag is either NA or Assert and Class 3 is NA. If the Class 1 tag is *Communication-management* (46%), then the Class 2 tag is *Expression* and the Class 3 tag is NA or Closing.

Table 2 shows the number of occurrences of all dialog acts in the training corpus for each of the eight classes. For all classes except **Information-level** and **Statement**,

#Occ	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8
541	CM	Exp	NA	NA	NA	NA	Bc	NA
383	CM	Exp	Cl	NA	NA	NA	NA	NA
329	Task	Ass	NA	NA	NA	NA	NA	NA
293	Task	NA	NA	Ad	NA	NA	NA	NA
244	Task	NA	NA	EIr	NA	NA	NA	NA
216	Task	Ass	NA	NA	Acc	R	NA	NA

Fig. 2. CM: Communication management; Ass: Assert; Exp: Expression; Cl: Closing; EIr: Explicit-Info-request; Ad: Action-directive; Acc: Accept; R: Response; Bc: Backchannel; NA: No applicable tag.

most (about 75%) of the time no tag is relevant (NA). Ignoring these the most frequent dialog acts are *Accept, Action-directive, Explicit-Information-request,* and *Back-channel*. It can be seen that many of the dialog acts occur only a few times. Since every utterance unit has an Information-Level tag (Class 1), the NA tag is never used.

5. Memory based learning

The goal of this work is to automatically detect the dialog acts associated with each utterance unit. A memory based learning methodology was adopted since such methods have been shown to be well adapted for natural language processing (Bosch et al., 2001; Daelemans et al., 1999) and work well with small amounts of data. The IB1-IG implementation of Machine Based Learning from the TiMBL software package (Daelemans et al., 2003) was employed using the Manhattan distance, one of the most basic metrics that works well with symbolic features. With this metric, 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 unknown one. 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. The model is formed by constructing the vectors for all of the training utterance units.

This work is based on the following hypotheses:

- The Dialog Act combinations are highly constrained as previously illustrated in Fig. 2.
- The initial words of the utterance unit are more important than the remaining words for identifying the dialog act for example "I'd like ...", "can you give me ..."
- Most of the information is encoded in specific entities.[2]

A question that arises here is what features are relevant for the vectors. We chose to use the number of utterance units in the turn, the words and the dialog act tags as features. Since our lexical hypothesis is that words in initial position are the most important, for the lexical features only the N first words (N_1 words) of each utterance units are used. For DA annotation, the previous tags in the turn are used as features.

The tag feature vector has eight items, one for each of the eight classes since each utterance unit can potentially receive one tag for each class. If none of the tags for a class is relevant, it is represented by NA (not applicable).

For DA annotation, a segmented speaker turn is input to the system which extracts the defined features and puts them into a vector, and determines the dialog acts:

$$W_1(\mathrm{UU}_1) = (w_1, w_2, \ldots, w_{N_1}) \quad (1)$$

$$\mathrm{DA}_1(\mathrm{UU}_1) = mbl[\#Utt., W_1] \quad (2)$$

where DA is a dialog act, UU is an utterance unit, w is a word and N_1 is the number of words used for the first utterance unit.

For instance, the turn of the Agent

U1: donnez-moi votre numéro de compte (*give me your account number*)

which has one utterance unit and the following dialog act tags:

Information-level = *Task*; **Influence-on-listener** = *Action-directive*

is represented by the following vector (with $N = 4$ words):

1 donnez-moi votre numéro *Task NA NA Action-directive NA NA NA NA*

The response of the Client:

U2: alors 256 132 34 56 7 (*then 256 132 34 56 7*)

is represented by the following vector (with $N = 4$ words):

1 alors 256 132 34 *Task Assert NA NA Accept true NA NA*.

Since our first hypothesis is that the succession of dialog acts is highly constrained, the classification is done in eight steps, one for each class. The classification of the vector (e.g. assigning a dialog act to it) is done by comparing the vector to all the examples in the training database. The result of this first classification is then considered as an element of the vector used to classify the next dialog act:

$$\mathrm{DA}_i(\mathrm{UU}_1) = mbl[\#Utts, W_1, \mathrm{DA}_1, \ldots, \mathrm{DA}_{i-1}] \quad (3)$$

The first utterance unit is thus tagged for all eight dialog act classes. The DA tags are determined in a sequential manner where tag_n depends on previous N tags. Different

[2] Specific entities can be though of as semantic classes and are introduced to reduce lexical variability and improve generalization.

Table 2
Number of occurrences of all dialog acts in the 3912 utterance units of the training corpus, grouped by class (1–8)

1 Information level	# Occ.
Com-mgt	1815
Out-of-topic	25
Task	1947
Task-mgt-Completion	21
Task-mgt-Order	30
Task-mgt-Summary	65
Task-mgt-SysCap	9
NA	0
2 Statement	**# Occ.**
Assert	1112
Commit	54
Explanation	73
Expression	1583
ReExplanation	3
Reassert	82
Re-Commit	2
NA	1003
3 Conventional	**# Occ.**
Closing	405
Opening	320
NA	3187
4 Influence-on-Listener	**# Occ.**
Action-directive	378
Expl-Confirm-request	114
Expl-Info-request	293
Impl-Confirm-request	26
Impl-Info-request	50
Offer	27
Open-Option	20
Re-Action-directive	10
Re-Confirm-request	3
Re-Info-request	3
Re-Offer	2
NA	2986
5 Agreement	**# Occ.**
Accept	536
Accept-part	5
Maybe	6
Reject	45
Reject-part	5
NA	3315
6 Answer	**# Occ.**
True	651
NA	3261
7 Understanding	**# Occ.**
Backchannel	616
Completion	29
Correction	13
Non-understanding	15
Repeat-rephrase	121
NA	3118
8 Communicative status	**# Occ.**
AbandStyle	5
AbandTrans	6
AbandChangeMind	8
AbandlossIdeas	7
Interrupted	39
Self-talk	6
NA	3841

experiments using different orders have been conducted. The best class order has been retained.

If there is more than one utterance unit in the turn, the first N_2 words of the next utterance unit are added to the vector containing the hypotheses for the previous utterance unit:

$$W_{j>1} = (w_1, w_2, \ldots, w_{N_2}) \tag{4}$$

$$\begin{aligned}&\mathrm{DA}_i(\mathrm{UU}_j)\\ &\quad = mbl[\#Utts, \mathrm{DA}_{1\ldots8}(\mathrm{UU}_1)\ldots\mathrm{DA}_{1\ldots8}(\mathrm{UU}_{j-1}),\\ &\qquad \mathrm{DA}_1\ldots\mathrm{DA}_{i-1}, W_1\ldots W_j]\end{aligned} \tag{5}$$

where W_j are the words for the utterance unit j if $j > 1$, UU_j is the utterance unit j, DA_i is the Dialog Act i and N_2 are the number of words added for utterance unit $j > 1$.

6. Experiments and results

In order to test the developed method, a first series of experiments were carried out using different configurations for the baseline models, and then experiments assessed the use of the dialog history, and evaluated language and task portability.

The first series of experiments were carried out using the GE_fr corpus (manually transcribed and segmented), with a model trained on the designated training portion. The division of the GE_fr corpus into two sets for training and testing purposes is shown in Table 3. Roughly one-third of the data has been reserved for test. Performance is reported in terms of DA tagging accuracy, since this is the most commonly used metric. Each tag is treated individually, so there are eight tags per utterance unit and a total of 13,688 tags in the GE_fr test data.

6.1. Results with the general model

Several experimental configurations were explored to test the general model. For the baseline configuration, different combinations of the number of words (N_1) for the first utterance unit in the turn have been explored, as well as different numbers of words added for each subsequent

Table 3
GE_fr training and test corpora

	Training	Test
# dialogs	94	40
# turns	2923	1350
# utterance units	3912	1711
# max of UU	6	4

Table 4
Dialog act detection rate on the GE_fr test data for different experimental setups using the General Model

	Entity condition			
	4 + 2 entities	4 + 4 entities	2 + 2 entities	2 + 4 entities
Baseline	84.5	84.7	84.2	84.4
Configuration 1	85.0	85.1	84.8	85
Configuration 2	85.6	85.7	85.4	85.6

Baseline: entities = words; Configuration 1: entities = specific entities; Configuration 2: entities = specific entities with separate Agent/Client models.

Table 5
Dialog Act success rate with Configuration 2 and the expected agreement $P(E)$ in percent for each class

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8
DA (%)	86.5	76.3	96.0	81.6	82.9	81.3	84.5	96.7
$P(E)$	45.3	32.9	67.5	59.6	68.5	73.1	62.3	94.1

Class 1: information level, Class 2: statement, Class 3: conventional, Class 4: influence-on-listener, Class 5: agreement, Class 6: answer, Class 7: understanding, Class 8: communicative status.

utterance unit (N_2). Results are reported here for the following setups:

- *2 + 2*: the first two words of each utterance unit
- *2 + 4*: the first two words of the first utterance unit and the first four words for each subsequent one
- *4 + 2*: the first four words of the first utterance unit and the first two words of each subsequent utterance unit
- *4 + 4*: the first four words of each utterance unit

A second set of experiments (Configuration 1) was designed to test our hypothesis that the information for dialog act annotation can be encoded by specific entities classes. The specific entities are:

- Named Entities which are expressions for people, places, organizations.
- Task Entities which are named entities that describe task or domain specific knowledge such as account number, account balance, transfer amount.
- Linguistic Entities which give structure to the utterances, for example "I'd like to ...".

These specific entities can be seen like the semantic clusters described in (Samuel et al., 1999). The main aim of defining specific entities is to reduce the lexical variability, and therefore reducing the model size and the search space and improving generalizability. All the data is automatically tagged with specific entities using rewrite rules which work like local grammars with specific dictionaries. The specific dictionaries were built using the semantically annotated training corpus described in (Amities Consortium, 2005, 2003). During processing all words matching specific entities are replaced by their respective classes.

For instance, the turn of the Agent:

U1: donnez -moi votre numéro de compte (*give me your account number*)

is replaced by the following one:

1 donnez -moi votre [accno]

The turn of the Client:

U2: alors 256 132 34 56 7 (*then 256 132 34 56 7*)

is replaced by the following one:

1 alors [num] [num] [num] [num] [num]

Because it seems obvious that the client and the agent speak differently, in Configuration two different models were constructed and evaluated for the two roles, one Client model and one Agent model.

Table 4 reports results for the three configurations using different entities conditions (the number of entities for N). For the baseline configuration which uses the original words in the transcript of what was said, the correct DA detection rate is about 84.7%, with the highest success obtained using $N = 4$ and an increment of four words.[3] The results are improved by about 0.5% when specific entities are used instead of words (Configuration 1), demonstrating the positive effect of reducing the lexical variety. The last entry (Configuration 2) gives results with separate Agent and Client models, which is seen to give an absolute gain of about 1% for all entities conditions.

The best result (85.7% accuracy) is obtained with the 4 + 4 specific entities setup and different models for the Agent and the Client. Table 5 compares the results of this best model with an expected agreement, i.e., the probability that the system chooses the correct dialog act by chance. In this case, $P(E)$ is simply the per-class sum of the square of the probability of each dialog act tag. This table shows the results for each of the eight classes. With the exception of Class 8, the models performs significantly better than an optimal random one.

[3] Going higher than four words did not improve results since there are not enough long utterance units.

6.2. Use of the dialog history information

Table 6 shows the dialog act detection rates obtained with the Configuration 2 model as a function of the utterance unit position. These results indicate that the dialog history, or more precisely the position of the utterance unit (a short history), has a different incidence and weight. In an attempt to improve the dialog act detection, a second series of experiments have been carried out making use of the dialog history. These experiments rely on two observations:

- There are links between the different utterance units in one turn and these links are structured.
- A dialog being a succession of turns, the dialog acts of one turn have an impact on the dialog acts of the next turn.

With the general model, all dialog acts of all of the previous utterance units in the turn are added as features to the current vector. For the following experiments, Eq. (5) was generalized as:

$$\mathrm{DA}_i(\mathrm{UU}_j) = mbl[\#Utt., \mathrm{DA}_1 \ldots \mathrm{DA}_{i-1}, W_j, C_j] \quad (6)$$

where the turn history information C_j is defined as[4]:

$$\begin{aligned} C(\mathrm{UU}_1) &= \emptyset \\ C(\mathrm{UU}_2) &= (W_1, DA(\mathrm{UU}_1)) \\ C(\mathrm{UU}_3) &= (W_1, W_2, DA(\mathrm{UU}_1), DA(\mathrm{UU}_2)) \\ C(\mathrm{UU}_4) &= (W_1 \ldots W_3, DA(\mathrm{UU}_1), DA(\mathrm{UU}_2), DA(\mathrm{UU}_3)) \end{aligned} \quad (7)$$

It can be seen in Table 6 that the results degrade when more than two utterance units are used (rows UU_3 and UU_4). This may be due to the inherent difficulties in classifying DAs in long turns, to changes of topic in long turns, to the lack of sufficient training data with more than two utterance units or to the noise introduced by the classification errors in the first two utterances. One corpus-driven hypothesis is that the tags of the previous utterance unit, which are added to the vector used to annotate the current utterance unit, are only helpful for tagging the second utterance unit in a turn.

In the first of a series of contrastive experiments, no turn history nor previous entities were used in annotating the third utterance unit. The contextual information used in this contrastive experiment is shown in Eq. (8), where the history is seen to be empty for the first and third utterance units. Table 7 repeats the results for the Configuration 2 model with the 4 + 4 entities setup which systematically incorporated all previous dialog acts of all previous UUs to determine the dialog acts of the current utterance, and also gives results when no turn history is used for the third utterance unit as shown by the following equation:

[4] Turns with more than four utterance units are extremely rare in both the training and test sets, so these were ignored.

Table 6
Dialog Act detection rate by utterance unit in GE_fr test data with Configuration 2 model (using Eq. (7))

# Utt.	4 + 2 entities	4 + 4 entities	2 + 2 entities	2 + 4 entities
UU_1	85.3	85.3	85.0	85.0
UU_2	87.6	88.4	87.4	88.7
UU_3	80.2	77.9	82.0	80.2
UU_4	79.2	79.2	79.2	79.2
Turn	85.6	85.7	85.4	85.6

For each utterance unit, the complete set of dialog acts (history) is used in determining its dialog acts.

Table 7
Dialog Act detection rate as a function of the number of utterance units (Configuration 2, 4 + 4 entities setup)

	UU_1	UU_2	UU_3	UU_4	Turn
Configuration 2	85.3	88.4	**77.9**	79.2	**85.7**
Exp 1	85.3	88.4	**85.3**	79.2	**85.9**

Exp 1: no turn history for the third utterance unit as shown in Eq. (8).

$$\begin{aligned} C(\mathrm{UU}_1) &= \emptyset \\ C(\mathrm{UU}_2) &= (W_1, DA(\mathrm{UU}_1)) \\ C(\mathrm{UU}_3) &= \emptyset \\ C(\mathrm{UU}_4) &= (W_3, DA(\mathrm{UU}_3)) \end{aligned} \quad (8)$$

Compared to the original Configuration 2 system, the dialog act detection rate for the third utterance unit is improved by about 7% and there is no change for utterance unit 4. It should be noted that in any case the results for utterance unit 4 are not very reliable since there are very few turns with 4 utterance units in the data (only 17 in the training and 6 in the test).

Concerning the second dialog history hypothesis, if the dialog acts of a turn have an incidence on the dialog acts of the next turn, then, it seems useful to capture a larger dialog history. This suggests adding the dialogic information of the last utterance unit of the previous turn to the first utterance unit of a turn, which can be expressed as:

$$\begin{aligned} C(\mathrm{UU}_1) &= (DA(\mathrm{last\_UU}(T_{n-1}))) \\ C(\mathrm{UU}_2) &= (W_1, DA(\mathrm{UU}_1)) \\ C(\mathrm{UU}_3) &= (W_1, W_2, DA(\mathrm{UU}_1), DA(\mathrm{UU}_2)) \\ C(\mathrm{UU}_4) &= (W_1 \ldots W_3, DA(\mathrm{UU}_1), DA(\mathrm{UU}_2), DA(\mathrm{UU}_3)) \end{aligned} \quad (9)$$

Results for the first and second utterance units are slightly improved using this history as shown in the first row of Table 8 Exp 2 (hypotheses).

The third contrastive experiment (Exp 3) combines the most successful conditions of the previous ones. For the first utterance unit the last utterance unit of the previous turn is used to provide history information. For the other utterance units, the previous utterance unit is used except for the third utterance unit for which no history information or previous entities are used. This is summarized by the following equations:

Table 8
Contrastive experiments with dialog history (Configuration 2, 4 + 4 entities setup)

	UU_1	UU_2	UU_3	UU_4	Turn
Configuration 2	85.3	88.4	**77.9**	79.2	**85.7**
Exp 2 (hypotheses)	**85.4**	**88.8**	77.9	79.2	**85.8**
Exp 2 (Oracle)	**88.1**	**89.0**	79.0	79.2	**88.0**
Exp 3 (hypotheses)	**85.5**	88.8	**85.3**	79.2	**86.1**
Exp 3 (Oracle)	**88.1**	89.0	**85.3**	79.2	**88.2**

Exp 2 and Exp 3. Oracle results use the reference DAs for the last utterance unit of the previous turn instead of the hypothesized DAs.

$$\begin{aligned} C(\mathrm{UU}_1) &= (DA(\mathrm{last\_UU}_(T_{n-1}))) \\ C(\mathrm{UU}_2) &= (W_1, DA(\mathrm{UU}_1)) \\ C(\mathrm{UU}_3) &= \emptyset \\ C(\mathrm{UU}_4) &= (W_3, DA(\mathrm{UU}_3)) \end{aligned} \tag{10}$$

As can be seen in Table 8, Exp 3 (hypotheses), results are improved for all utterance units and this model gives the best overall results and for each utterance unit. Looking closely at the data we observed that, when a turn contains more than 2 units, the speaker usually changes their intention after the second unit. However, there is still some room for improvement as can be seen by the Oracle results in Table 8 where the reference DAs for the last utterance unit of the previous turn are used instead of the hypothesized DAs. The Oracle detection rate is about 2% higher for the first utterance unit, which gives an indication the influence of errors in the detected dialog acts of the second and third utterance units (the most probable last utterance units of the previous turn) on the detection of the next dialog acts.

6.3. Cross-domain and cross-language conditions

In order to further test our second hypothesis about the role of lexical information, models trained on the GE_fr corpus were applied to the CAP_fr corpus (a change of task) and to the GE_eng corpus (a change of language).

Table 9
Dialog act detection success rate on CAP_fr test data for different experimental setups with GE_fr general models

	4 + 2 entities	4 + 4 entities	2 + 2 entities	2 + 4 entities
Baseline	76.8	76.3	76.3	75.7
Configuration 1	75.9	76.3	76.3	76.8
Configuration 2	77.7	77.7	77.1	77.2

The baseline configuration uses words, Configuration 1 model uses specific entities and Configuration 2 separate Agent/Client models.

The GE_fr model can be applied to the other corpora since it is used after specific entity tagging. The French and English taggers produce the same tag sets. The results obtained on the CAP_fr corpus are presented in the Table 9 using three configurations of the general model. As for the GE_fr data, the best dialog act detection success (about 77%) is obtained with the Configuration 2 model. The results obtained on the GE_eng corpus are presented in the Table 11. The best results are once again obtained with the Configuration 2 model, with a correct detection rate of about 75%. Table 10 compares the results of the best model with an expected agreement, i.e. the probability that the system chooses the correct dialog act by chance. With the exception of Class 7 for GE_eng data, the results of our models are significantly better than an optimal random one.

The dialog history was added to the general model in a manner analogous to the previous experiments (Exp 1–Exp 3). To asses the performance under cross-domain and a cross-language conditions experiments were carried out with the 2 + 2 entities setup. As shown in Table 12 (hypotheses entries) for CAP_fr test, all three experiments incorporating dialog history improve performance over the general Configuration 2 model. This is not the case for the GE_eng test data, where, as shown in Table 13, the three experiments including the dialog history worsen the dialog act detection performance.

Overall the results are seen to be worse, which we attribute to the errors in both specific entity (about 70%-80% for the English tagger versus 90% for the French tagger)

Table 10
Dialog Act success rate with Configuration 2 and the expected agreement $P(E)$ in percent for each class

	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8
CAP_fr DA(%)	83.4	60.2	84.1	75.6	79.4	77.4	72.3	92.7
$P(E)$	34.6	31.6	78.6	48.3	77.9	75.6	68.2	92.1
GE_eng DA(%)	70.7	40.6	92.3	59	72.6	81.2	74.7	99.7
$P(E)$	54.8	37.5	81.4	46.3	32.3	56.3	76.1	99.4

Class 1: information level, Class 2: statement, Class 3: conventional, Class 4: influence-on-listener, Class 5: agreement, Class 6: answer, Class 7: understanding, Class 8: communicative status.

Table 11
Dialog act detection success rate on GE_eng test data for different experimental setups with GE_fr general models

	4 + 2 entities	4 + 4 entities	2 + 2 entities	2 + 4 entities
Configuration 1	70.2	70.0	74.0	73.9
Configuration 2	69.9	70.0	74.8	74.7

The Configuration 1 model uses specific entities and Configuration 2 specific entities with separate Agent/Client models.

Table 12
Success rate for DA detection on CAP_fr test data with GE_fr models and different ways of incorporating dialog history information

	UU_1	UU_2	UU_3	UU_4	Turn
Configuration 2 (2 + 2 words)	77.0	79.7	69.7	72.5	77.1
Exp 1	77.0	79.7	79.8	75.0	77.4
Exp 2 (hypotheses)	77.1	80.2	72.1	75.0	77.3
Exp 2 (Oracle)	80.4	80.3	72.1	72.5	80.2
Exp 3 (hypotheses)	77.2	80.5	79.8	75.0	77.7
Exp 3 (Oracle)	80.4	80.3	79.8	75.0	80.3

Table 13
Success rate for DA detection on GE_eng test data with GE_fr models and different ways of incorporating dialog history information

	UU_1	UU_2	UU_3	UU_4	Turn
Configuration 2 (2 + 2 words)	74.7	73.8	78.6	75.0	74.8
Exp 1	74.7	73.8	78.6	64.3	74.7
Exp 2 (hypotheses)	63.9	74.5	68.1	67.9	65.3
Exp 2 (Oracle)	81.3	77.4	68.6	73.2	80.3
Exp 3 (hypotheses)	65.6	76.2	78.6	64.3	67.6
Exp 3 (Oracle)	81.3	77.4	78.6	64.3	80.7

and dialog act tags of the last utterance units, which are used as features in order to predict the DAs in first utterance units in next turn. This observation is supported by the Oracle results shown in Tables 12 and 13 for the CAP_fr and GE_eng corpora respectively. If the correct DA tags are used to provide the dialog history, then the inclusion of this information systematically improves performance for CAP_fr and almost always for GE_eng.

These cross task and cross-language experiments used the Configuration 2 model with the 2 + 2 entities setup. In order to verify that this setup was appropriate, the 4 + 4 and 4 + 2 setups were tested on CAP_fr for Exp 3. Both gave a correct DA detection rate of 76.8%, which is lower than the 77.8% obtained with the 2 + 2 entities setup.

Fig. 3 summarizes the results of the above experiments. The figure illustrates that the general dialogic structure has been captured in the model. From the Oracle results it can be seen this model is potentially good under cross-language and cross domain conditions, and in practice the model is demonstrated to perform reasonably well.

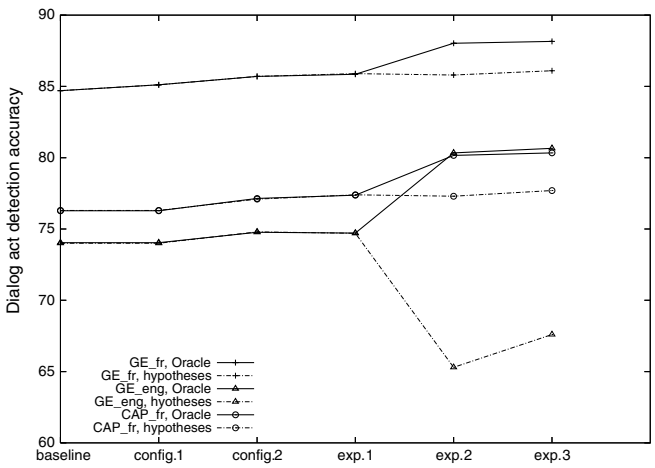

Fig. 3. Summary of the dialog act detection experiments.

7. Analyses and discussion

This section analyzes the experimental results and discusses them according to the the different dialog act classes.

7.1. Classes and dialog act detection error rate

As stated previously, the experiments making use of the dialog act history are based on two hypotheses, that the links between the different utterance units in a turn are structured and that a dialog being a succession of turns, dialog acts of a turn impact the dialog acts of the next turn. If these hypotheses are correct, then the dialog acts referring to previous information should have a better detection rate with models including the dialog history.

Four of the dialog act classes contain tags that refer to previous turns or utterance units. There are: **Influence-on-Listener**, **Agreement**, **Answer** and **Understanding**. Table 14 gives the DA detection rates for these with and without the inclusion of history information for the GE_fr test corpus. There is an improvement of about 1% for **Influence-on-listener**, **Agreement**, and **Understanding**, and of about 2% for **Answer** with the third method compared to the Configuration 2 model. Similar observations can be made for the CAP_fr corpus as shown in (Table 15). However the results on the GE_eng corpus are mostly worse than the Configuration 2 model. For both of these corpora a much larger improvement is obtained when the Oracle history is used. These results support our hypotheses on the global structure of the dialog, and how the dialog history can capture the reactions of one participant to prior actions of the other.

7.2. Error analysis

While all errors are equal for the evaluation metric, in the context of using the results of DA detection for dialog management different kinds of errors will have different impacts. In particular, confusions between the *Opening* and *Closing* **Conventional** tags, and between the different kinds of **Agreement** are important.

It can be seen in Table 16 that the vast majority of errors on the *Opening* or *Closing* tag are not confusions between

Table 14
Dialog Act success rate by class for the GE_fr corpus

	Influence on listener	Agreement	Answer	Understanding
Conf. 2	82.9	83.6	81.6	84.3
Exp 3 (hyp.)	83.9	84.7	83.8	85.4
Exp 3 (Oracle)	84.9	88.3	88.4	88.5

Table 15
Dialog act success rate and classes on the GE_eng and CAP_fr corpora

	Models	Influence on listener	Agreement	Answer	Understanding
CAP_fr	Conf. 2	73.1	79.0	77.5	74.5
	Exp 3 (hyp.)	74.2	81.0	80.1	75.9
	Exp 3 (Oracle)	75.5	85.9	85.4	79.0
GE_eng	Conf. 2	59.0	72.6	81.2	74.7
	Exp 3 (hyp.)	56.8	78.6	67.8	71.4
	Exp 3 (Oracle)	63.5	84.5	89.2	71.0

Table 16
Confusion matrix for **Conventional** tags (GE_fr test data)

Hyp.	Ref.			Total hyp.
	NA	*Closing*	*Opening*	
NA	1346	38	9	1393
Closing	40	149	2	191
Opening	4	1	122	127
Total ref.	1390	188	133	1711

Table 17
Confusion matrix for **Agreement** tags (GE_fr test data)

Hyp.	Ref.				Total hyp.
	NA	*Accept*	*Reject*	*Maybe*	
NA	1280	121	22	2	1425
Accept	94	167	7	1	269
Reject	11	4	2	0	17
Maybe	0	0	0	0	0
Total ref.	1385	292	31	3	1711

these two tags but between NA and the concerned tag. Of the 191 hypothesized *Closing* tags, 40 were NA and 2 opening in the reference annotations. Table 17 shows a confusion matrix for the **Agreement** tags. It can be seen that the confusion is more often between NA and the concerned tag than between two dialog acts. There were 102 hypothesized *Accept* tags that were not in the references. Of these 94 were manually tagged as NA, 7 as *Reject* and 1 as *Maybe*. We consider that confusions between NA and a dialog act are less important than between two different dialog acts, especially if these two acts are contradictory.

7.3. Automatic dialog act detection and inter-annotator agreement

A study on inter-annotator agreement (IAg) on a set of 60 dialogs from the GE_fr corpus was reported in (Hardy et al., 2003). In this section, these human–human IAg results are compared to the IAg between one human annotator and our automatic system. Inter-annotator agreement was measured according to the kappa statistic Carletta (1996):

$$k = \frac{P(A) - P(E)}{1 - P(E)} \tag{11}$$

where $P(A)$ is the proportion of times that the annotators agree and $P(E)$ is the proportion of times that we would expect the annotators to agree by chance. If there is complete agreement among the annotators, then $K = 1$; whereas if there is no agreement other than what would be expected by chance, then $K = 0$.

Fig. 4 plots the kappa value for both the human–human IAg and human-system IAg. The kappa values are seen to be slightly higher for the human-system IAg than for the Human–human IAg for most of the classes. For **Class 1:** (Information level) the kappa value is a little better for the human–human IAg than for the human-system IAg. For **Class 4:** (Influence-on-listener) and **Class 6:** (Answer), the human–human IAg is better than for the human-system IAg. These classes are the ones that require the largest dialog context. Classes 5 and 8 have a higher system-human kappa value than human–human ones. These two classes contain tags found to be more confusing for the human annotators. For instance, the distinction between Accept and Accept-part in **Class 5:** (Agreement) is very subjective. The system seems to have converged to the most common choices between the different annotators.

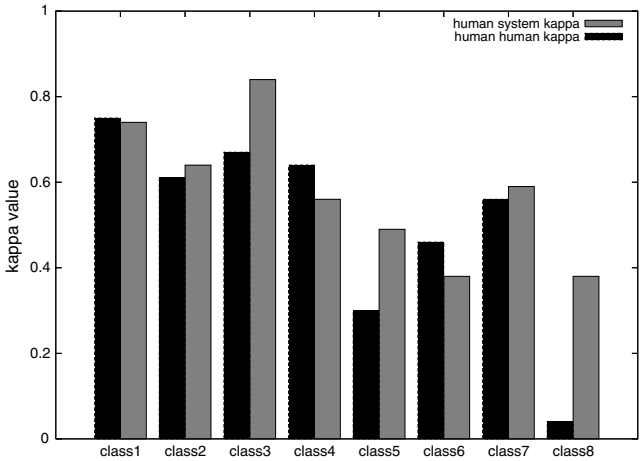

Fig. 4. Kappa value for human–human IAg and human-system IAg.

8. Conclusions and perspectives

This paper has reported on research aimed at automatic dialog act tagging for different corpora. Starting with the AMITIES multi-level dialog annotations based on DAMSL, a set of eight 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 corpus. The basic system uses a feature set for each speaker turn comprised of the number of utterance units in the turn, the previous hypothesized dialog acts and N entities per utterance unit. Data normalization using specific entities was explored in order to reduce the language and task dependency of the approach, as well as different models for the Agent and the Client. With the basic model (using word-based features), on the GE_fr corpus, the dialog act detection rate was about 84.7%, which increases to 85.1% when specific entities are included in the word-based model and to 85.7% when different speaker models (agent/client) are used.

A study of the data showed that there are strong constraints among the dialog act features in a single utterance unit, as well as between successive turns. After an analysis of the results using the basic system, some dialog history information was added to the feature vector. This information attempts to capture the observed dialog structure. The best model incorporating the within turn dialog history, where the history is ignored for the third utterance, has a correct dialog act detection rate of 86.1%.

Using the same model under cross-domain conditions results in a dialog act detection rate of 77.7% on the CAP_fr data. Under cross-language condition the model without dialog history obtains the best dialog act detection rate of 74.8%. The model using history information gave worse results. Our interpretation is that the dialog act detection error rate of the last utterance units are not good enough to be used in such a model, particularly given that the Oracle results show an improvement.

These results support our underlying hypotheses that most of the information is encoded in specific entities and that the dialog structure is important information for predicting dialog acts. Experimental results reported with an Oracle method lend additional support to these hypotheses.

In order to assess how well the dialog acts are being detected, the system agreement with the human annotator was compared to previously reported results on human–human inter-annotator agreement. This comparison naturally leads to the question of how reliable are humans at annotating dialog acts, and what is the most appropriate manner to evaluate the annotations.

The experiments and results reported in this paper assumed that the number of utterance units and the location 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 (Rosset and Lamel, 2004) showed that using a simple 4-gram language model could accurately predict the number of utterance units but that the localization of the boundary was less good. Our first result with a completely automatic dialog act annotation system gives a 77% dialog act detection rate with a Configuration 2 model (compared to 85% with known boundaries). It is likely that other sources of information of an acoustic nature, such as prosodic information (Ang et al., 2005; Ji and Bilmes, 2005) could be of use for predicting the boundary localizations and the dialog acts.

Acknowledgements

This research was partially financed by the European Commission under the projects FP5 IST-2000-25033 Amitiés and FP6 Integrated Project 506909 CHIL.

References

Allen, J., Core, M., 1997. Draft of DAMSL: Dialog Act Markup in Several Layers, October, http://www.cs.rochester.edu/research/cisd/resources/damsl.

Amities Consortium. AMITIES final report. June 2005.

Anderson, A., Bader, M., Bard, E., Doherty, G.M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H.S., Weinert, R., 1991. The HCRC Map Task Corpus. Language and Speech 34, 351–366.

Ang, J., Liu, Y., Shriberg, E., 2005. Automatic dialog act segmentation and classification in multiparty meetings. In: ICASSP'05, Philadelphia, April I, pp. 1061–1064.

Barras, C., Geoffrois, E., Wu, Z., Liberman, M., 2001. Transcriber: development and use of a tool for assisting speech corpora production. Speech Communication 33 (1–2), 5–22.

Bonneau-Maynard, H., Rosset, S., 2003. A semantic representation for spoken dialogs. In: ISCA Eurospeech'03, Geneva, September, vol. 1, pp. 253–256.

van den Bosch, A., Krahmer, E., Swerts, M., 2001. Detecting problematic turns in human–machine interactions: rule-induction versus memory-based learning approaches. In: ACL'00, New Brunswick, pp. 499–606.

Brill, E., 1995. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. Computational Linguistics 21 (4), 543–566.

Carletta, J., 1996. Assessing agreement on classification tasks: the kappa statistic. Comput. Linguist. 22 (2), 249–254.

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

Chu-Carroll, J., A statistical model for discourse act recognition in dialogue interactions. In: Applying Machine Learning to Discourse Processing. Papers from the 1998 AAAI Spring Symposium. Technical Report SS-98-01, AAAI Press, Menlo Park, CA, pp. 12–17.

Daelemans, W., van den Bosch, A., Zavrel, J., 1999. Forgetting exceptions is harmful in language learning. Machine Learn. 34, 11–43.

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

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

Ehara, T., Ogura, K., Morimoto, T., 1990. ATR Dialogue Database. ICSLP'90, Kobe, Japan, pp. 1093–1096.

Hardy, H., Baker, K., Bonneau-Maynard, H., Devillers, L., Rosset, S., Strzalkowski, T.T., 2003. Semantic and dialogic annotation for automated multilingual customer service. In: ISCA Eurospeech'03, Geneva, September, 1, pp. 201–204.

Hirschberg, J., Litman, D.J., 1993. Empirical studies on the disambiguation of Cue phrases. Comput. Linguist. 19 (3), 501–530.

Isard, A., Carletta, J.C., 1995. Replicability of transaction and action coding in the map task corpus. In: AAAI Spring Symposium: Empirical Methods in Discourse Interpretation and Generation, pp. 60–67.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, C., 2003. The ICSI meeting corpus, In: ICASSP'03, Hong Kong.

Jekat, S., Klein, A., Maier, E., Maleck, I., Mast, M., Quantz, J.J., 1995. Dialog Acts in Verbmobil. Verbmobil Report 65, Universitat Hamburg, DFKI Saarbrucken, Universitat Erlangen, TU Berlin.

Ji, G., Bilmes, J., 2005. Dialog Act tagging using graphical models. In: ICASSP 2005, Philadelphia, April 2005.

Jurafsky, D., Shriberg, E., Biasca, D., Switchboard-DAMSL labeling project coder's manual. Technical report 97-02, University of Colorado, Institute of Cognitive Science, Boulder, http://www.stanford.edu/jurafsky/ws97/manual.august1.html.

Nagata, M., 1992. Using pragmatics to rule out recognition errors in cooperative task-oriented dialogues. In: ICSLP'92, Banff, Canada, October.

Reithinger, N., Klesen, M., 1997. Dialogue act classification using language models. Eurospeech'97, Rhodes, pp. 2235–2238.

Reithinger, N., Engel, R., Kipp, M., Klesen, M., 1996. Predicting dialog acts for a speech to speech translation system. In: ICSLP 1996, Philadelphia, October.

Rosset, S., Lamel, L., 2004. Automatic detection of dialog acts based on multi-level information. In: ICSLP'04, Jeju Island, October, pp. 540–543.

Samuel, K., Carberry, S., Vijay-Shanker, K., 1998. Dialogue act tagging with transformation-based learning. COLING-ACL, 1150–1156.

Samuel, K., Carberry, S., Vijay-Shanker, K., 1999. Automatically selecting useful phrases for dialogue act tagging. In: Proceedings of the Fourth Conference of the Pacific Association for Computational Linguistics, Waterloo, Ontario, Canada.

SRI Transcripts. 1992. Transcripts derived from audiotape conversations made at SRI International. Menlo Park, CA, Prepared by Jacqueline Kowtko under the direction of Patti Price.

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

Surendran, D., Levow, G.-A., 2006. Dialog act tagging with support vector machines and hidden markov models. In: Interspeech'06, Pittsburgh, PA, September.

Traum, D., 2000. 20 questions on dialog act taxonomies. Journal of Semantics 17 (1), 7–30.

Webb, N., Hepple, M., Wilks, Y., 2005. Dialog act classification based on intra-utterance features. In: Proceedings of the AAAI Workshop on Spoken Language Understanding.