AMI Consortium

http://www.amiproject.org/

Funded under the EU Sixth Framework Programme

Multimodal interfaces action line of the IST Programme

Integrated Projects

AMI (IST-506811) and AMIDA (IST–033812)

State of the Art Report

Automatic Dialogue Act Recognition

November 6, 2007

AMI Consortium State of the Art Report

Automatic Dialogue Act Recognition

November 6, 2007

Abstract

A DA is a construct that describes the role that an utterance plays in a conversation and provides a bridge between an orthographic (word-level) transcription, and a richer representation of the discourse. The reliable recognition of the DA sequence in a conversation, and the resulting knowledge of the discourse structure, plays an important role in the development of applications such as: action items detection, decision detection, automatic summarisation, topic segmentation, dialogue structure annotation, etc. DA recognition systems are usually based on supervised statistical approaches: a model is learned (trained) from a set of annotated examples, and then evaluated on unseen data. This process requires to collect and manually annotate relevant amounts of conversational data. Therefore several annotated corpora have been produced in the last decade, giving birth to multiple DA annotation schemes. The DA recognition task comprises two related sub-tasks: segmentation, and classification or tagging. DA segmentation consists of subdividing the conversation into unlabelled DA segments closer to those manually annotated. Unlabelled DA segments are then classified and tagged with the most likely DA label. These tasks may be performed concurrently (joint DA recognition) or sequentially. Multiple evaluation metrics have been proposed for segmentation, classification and the DA recognition task.

1 Introduction

The concept of dialogue acts (DAs) is based on the speech acts described by Austin [1962] and by Searle [1969]. The idea is that speaking is acting on several levels, from the mere production of sound, over the expression of propositional content to the expression of the speaker's intention and the desired influence on the listener. Dialogue acts are labels for utterances which roughly categorise the speaker's intention.

As such, they are useful for various purposes in a dialogue or meeting processing situation. For example DAs can be used as elements in a structural model of a meeting. A simple example would be a browser which highlights all points where a suggestion or offer was recognised. Often DA labels serve also as elementary units to recognise higher levels of structure in a discourse. DAs may also control the processing of discourse content. To generate abstractive summaries, for example, content is extracted from utterances, and integrated in a discourse memory depending on the DAs of the utterances.

The dialogue act recognition process consists of two subtasks: segmentation and classification (tagging). The first step is to subdivide the sequence of transcribed words in terms of DA segments. The goal is to segment the text into utterances that have approximately similar temporal boundaries to the annotated DA units. The second step is to classify each segment as one of the DA classes from the adopted DA annotation scheme. These two steps may be performed either sequentially (segmentation followed by classification) or

jointly (both tasks carried out simultaneously by an integrated system). Although most of the work on automatic DA processing have been focused on the tagging task, assuming knowledge of the reference DA segmentation; novel integrated DA recognition frameworks are growing in popularity.

2 Dialogue Act Annotated Data Resources

Any effort to recognize dialogue acts requires data. The usual practice is to employ supervised machine learning, using material that has been hand-transcribed and then hand-annotated with a suitable dialogue act scheme. Since creating this sort of data is expensive, most efforts re-use an existing data set, or corpus, wherever they can. There are a number of factors that need to be balanced in deciding on a corpus:

- how much data is available, since more data usually implies better results.
- how many dialogue act classes the scheme contains. Classifiers have trouble learning too many distinctions.
- how well distributed the classes are. Classifiers work best with relatively equal numbers of examples for the various classes.
- how reliable the hand-annotation is. If there are several human annotators involved and they tend to assign different classes for the same kind of material, the inconsistency makes it difficult both for the classifier and for evaluating the results. On the other hand, if there is only one annotator, but no one else would assign classes the same way, what the classifier is learning may not be useful.
- what language the data is for . The vast majority of available material and published work is on English.
- whether the data is generally available or access is restricted in some way.

For researchers who are interested in dialogue act recognition as an end in itself, for instance, as a means of trying out various machine learning algorithms, these are the primary considerations. However, where the recognizer is being built for use in an end application, it is also important that the dialogue act scheme makes the distinctions that the end application actually needs. It is no use having an extremely accurate classifier that cannot identify "backchannel" utterances such as "mhm-hmm" in a system that requires a very natural style of interaction, for instance. In addition, it is important that the material to which a scheme has been applied be similar enough to what the end application will encounter for what the classifier learns to transfer well. The closer the material, the better, which is why systems developers almost always collect at least some human-human dialogues that are as close to what the system will do as they can get. Since most applications involve having a system perform a task for the user, such as booking travel, task-oriented data is of the most use, but simply making the distinction between task-oriented dialogues and more free-ranging conversations is not enough. Differences such as using non-native or elderly speakers can have a large effect where these are the target users for the end application. Similarly, whether or not speakers use telephones changes their behaviour.

In dialogue act recognition, it is not necessary to learn every label from the hand-annotated scheme. Hand-annotated data can be transformed into a smaller set of labels by grouping individual classes together. This sort of transformation is sometimes called a "classmap".

Because of this practice, recent dialogue act scheme designers often include more labels than a classifier can learn, and then perform an analysis of the hand-annotated data once it is complete in order to decide what transform to use. There can be several acceptable classmaps for the same scheme, depending on how the resulting classifier is to be used. In constructing classmaps, it is common to put together classes that the human annotators frequently confuse with each other to make the data more consistent. However, it is important to ensure the label groupings also make sense in terms of the end application. The smallest schemes tend to provide 12-15 mutually exclusive labels and expect only a few, or none, to be combined.

Dialogue act schemes also differ in how they instruct the human annotator to segment the dialogue. For some schemes, the segmentation is purely ideational; annotators are to decide on segmentation by breaking the material into pieces that each express a complete meaning. For others, the segmentation is partly mechanical – for instance, the annotator may be instructed to provide segment boundaries at long pauses. Occasionally, the scheme assumes that the material has been presegmented by completely mechanical means (e.g., as in the Japanese Map Task Corpus, [Horiuchi et al., 1999]). Not surprisingly, ideational segmentations show the most disagreement among the human annotators, but they also in theory supply the most information.

Finally, dialogue act schemes vary in whether they adhere to dialogue act theory in simply segmenting and labelling material based on speaker intentions, or whether they include labels that are, strictly speaking, not dialogue acts at all. The most common addition is labels that identify disfluent material, particularly at the beginning of turns. The reason for their inclusion is to improve the results of language modelling on the data. Theoretically, the disfluent material belongs within an adjacent act, but because questions typically have a different syntactic form from statements and commands, the words that occur at the beginning of an act are important for determining what the act is. Dialogue act schemes that include these quasi-acts assume they will be used to strip this material out, sometimes as a first step before proper dialogue act recognition.

Klein et al. [1998] provides a detailed survey of early dialogue act schemes. The corpora currently in most common use for dialogue act recognition are the following:

The HCRC Map Task Corpus [Anderson et al., 1991], using a scheme developed for it [Carletta et al., 1997]. The scheme is intended to be general, but the material coded involves two people navigating around a simple map. One unusual aspect of this material is that it contains higher level dialogue structure coding. Many corpus users consider the fact that the speakers are Scottish as a disadvantage. It is also relatively small: 128 dialogues resulting in around 10 hours of speech.

The related DCIEM Map Task Corpus [Bard et al., 1996], which replicates the same task but using Canadian army reservists and includes sleep deprivation conditions comparing the effects of various drugs.

The Switchboard Corpus [Godfrey et al., 1992] using SWBD-DAMSL [Jurafsky et al., 1997b]. The underlying material consists of telephone conversations on a fixed set of topics, resulting in more than 200000 utterances and 1.4 millions transcribed words. The SWBD-DAMSL annotation scheme comprises 226 unique tags, which were subsequently clustered into 42 broad DA classes. A common concern when using this material is that that a very large proportion of the dialogue acts are of the

same type (basic statements).

The ICSI Meeting Corpus [Janin et al., 2003], using the ICSI-MRDA scheme [Shriberg et al., April-May 2004]. The corpus contains audio recordings of research group meetings. The scheme requires each act to be labelled along a number of semi-orthogonal dimensions, with thousands of tag combinations that are theoretically possible.

The AMI Meeting Corpus [Carletta, In Press], using a scheme developed for it. This corpus is unusual in involving non-native speakers of English and in making available a range of videos and other outputs that capture behaviour more fully than usual. The dialogue act scheme includes some extra features related to acts, such as information about addressing and some very rudimentary discourse structure.

2.1 The ICSI Meeting Corpus and Dialogue Act Tag Set

The ICSI meetings corpus [Janin et al., 2003] consists of 75 naturally occurring research group meetings at the International Computer Science Institute in Berkeley during the years 2000–2002, and recorded using close-talking microphones worn by each participant (in addition, there were also four tabletop microphones). Each meeting lasts about one hour and involves an average of six participants, resulting in about 72 hours of multichannel audio data. The corpus contains human-to-human interactions recorded from naturally occurring meetings. Moreover, having different meeting topics and meeting types, the data set is heterogeneous both in terms of content and structure.

Orthographic transcriptions are available for the entire corpus, and each meeting has been manually segmented and annotated in terms of Dialogue Acts, using the ICSI MRDA scheme [Shriberg et al., April-May 2004]. The MRDA scheme, outlined in table 1, is based on a hierarchy of DA types and sub-types (11 generic tags and 40 specific sub-tags), and allows multiple sub-categorisations for a single DA unit. A DA is usually composed by a single generic tag (statement, question, etc.) and several specific sub tags. This extremely rich annotation scheme results in more than a thousand unique DAs, although many are observed infrequently. To reduce the number of sparsely observed categories, a reduced set of five broad DA categories has been defined in [Ang et al., 2005, Zimmermann et al., 2006a]. Unique DAs were manually grouped into five generic categories: statements, questions, backchannels, fillers and disruptions. The distribution of these categories across the corpus is shown in table 2. Note that statements are the most frequently occurring unit, and also the longest, having an average length of 2.3 seconds (9 words). All the other categories (except backchannels which usually last only a tenth of a second) share an average length of 1.6 seconds (6 words). An average meeting contains about 1500 DA units.

In order to have directly comparable results a formal subdivision has been proposed by Ang et al. [2005]: a training set of 51 meetings (about 80.000 DAs), 11 meetings for the development task (13.500 DAs), and a test set composed by 11 meetings and 15.000 DAs. This leaves out 2 of the 75 meetings, which were excluded because of their different nature.

Statement	
s	**Statement**

Questions	
qy	**Yes/No Question**
qw	**Wh-Question**
qr	**Or Question**
qrr	**Or Clause After Y/N Question**
qo	**Open-ended Question**
qh	**Rhetorical Question**

Floor Management	
fg	**Floor Grabber**
fh	**Floor Holder**
h	**Hold**

Backchannels	
b	**Backchannel**
bk	Acknowledgement
ba	Assessment/Appreciation
bh	Rhetorical Question Backchannel

Responses	
aa	Accept
aap	Partial Accept
na	Affirmative Answer
ar	Reject
arp	Partial Reject
nd	Dispreferred Answer
ng	Negative Answer
am	Maybe
no	No Knowledge

Action Motivators	
co	Command
cs	Suggestion
cc	Commitment

Checks	
f	Follow Me
br	Repetition Request
bu	Understanding Check

Restated Information	
r	Repeat
m	Mimic
bs	Summary
bc	Correct Misspeaking
bsc	Self-Correct Misspeaking

Supportive Functions	
df	Defending/Explanation
e	Elaboration
2	Collaborative Completion

Politeness Mechanisms	
bd	Downplayer
by	Sympathy
fa	Apology
ft	Thanks
fw	Welcome

Further Descriptions	
fe	Exclamation
t	About-Task
tc	Topic Change
j	Joke
t1	Self Talk
t3	Third Party Talk
d	Declarative Question
g	Tag Question
rt	Rising Tone

Disruptions	
%	*Indecipherable*
%-	*Interrupted*
%–	*Abandoned*
x	*Nonspeech*

Nonlabeled	
z	Nonlabeled

Table 1: DA labels used for the annotation of the ICSI meeting corpus: **generic tags**, specific tags and *disruptions*.

Dialogue Act	% of total DA units	% of corpus length
Statement	58.2	74.5
Disruption	12.9	10.1
Backchannel	12.3	0.9
Filler	10.3	8.7
Question	6.2	5.8

Table 2: Distribution of DAs by % of the total number of DA units and by % of corpus length.

2.2 The AMI Dialogue Act Tag Set

The AMI meeting corpus [Carletta et al., 2005] is a multimodal collection of annotated meeting recordings. It consists of about 100 hours of meetings collected in three instrumented meeting rooms. About two thirds of the corpus consists of meetings elicited using a scenario in which four meeting participants, playing different roles on a team, take a product development project from beginning to completion. The scenario portion of the corpus consists of a number of meeting series, with four meeting per series. Each series of four meetings involves the same four participant roles, and comprises project kick-off, functional design, conceptual design, and detailed design meetings. The aim of the corpus collection was to obtain a multimodal record of the complete communicative interaction between the meeting participants. To this end, the meeting rooms were instrumented with a set of synchronised recording devices, including lapel and headset microphones for each participant, an 8-element circular microphone array, six video cameras (four close-up and two room-view), capture devices for the whiteboard and data projector, and digital pens to capture the handwritten notes of each participant. The corpus has been manually annotated at several levels, including orthographic transcriptions, various linguistic phenomena including head and hand movements, and focus of attention[1]. Most of the scenario data in the AMI corpus, over 100,000 utterances, have been annotated for dialogue acts. The DA annotation scheme for the AMI corpus [2] , outlined in table 3, is based around a categorization tailored for group decision making, and consists of 15 dialogue act types (table 3), which are organised in six major groups:

- Information exchange: giving and eliciting information
- Possible actions: making or eliciting suggestions or offers
- Commenting on the discussion: making or eliciting assessments and comments about understanding
- Social acts: expressing positive or negative feelings towards individuals or the group
- Other: a remainder class for utterances which convey an intention, but do not fit into the four previous categories
- Backchannel, Stall and Fragment: classes for utterances without content, which allow complete segmentation of the material

Each DA unit is assigned to a single class, corresponding to the speaker's intent for the utterance. The distribution of the DA classes, shown in table 3, is rather imbalanced,

1. The annotated corpus is freely available from `http://corpus.amiproject.org`
2. Guidelines for Dialogue Act and Addressee Annotation V1.0, Oct 13, 2005. `http://mmm.idiap.ch/private/ami/annotation/dialogue_acts_manual_1.0.pdf`

Group	Dialogue Act		Frequency	
Segmentation	fra	Fragment	14348	14.0%
	bck	Backchannel	11251	11.0%
	stl	Stall	6933	6.8%
Information	inf	Inform	28891	28.3%
	el.inf	Elicit Inform	3703	3.6%
Actions	sug	Suggest	8114	7.9%
	off	Offer	1288	1.3%
	el.sug	Elicit Offer or Suggestion	602	0.6%
Discussion	ass	Assessment	19020	18.6%
	und	Comment about Understanding	1931	1.9%
	el.ass	Elicit Assessment	1942	1.9%
	el.und	Elicit Comment about Understanding	169	0.2%
Social	be.pos	Be Positive	1936	1.9%
	be.neg	Be Negative	77	0.1%
Other	oth	Other	1993	2.0%
Total			102198	100.0%

Table 3: The AMI Dialogue act scheme, and the DA distribution in the annotated scenario meetings.

with over 60% of DAs corresponding to one of the three most frequent classes (inform, backchannel or assess). Over half the DA classes account for less than 10% of the observed DAs. This annotation scheme is different to the one used for the ICSI corpus (section 2.1), thus it is not possible to test a DA recognition system developed on the AMI data on the ICSI corpus or vice-versa.

The scenario meetings are organised in 35 series of (normally) four meetings, which have been split into designated training, development and evaluation sets. 25 series of meetings have been assigned to the training set, five to the development and five to the test set (table 4). For the purpose of cross-validation, a split into ten parts was also defined ; being this split useful both for ten-fold and five-fold cross-validation.

Subset	Meetings	#meetings	#series
Training set	ES2002, ES2005-2010, ES2012-2016 IS1000-1007 TS3005 TS3008-3012	98	25
Development set	ES2003, ES2011, IS1008, TS3004, TS3006	20	5
Evaluation set	ES2004, ES2014, IS1009, TS3003, TS3007	20	5
All scenario data		138	35

Table 4: The split of the AMI scenario data into training, development and evaluation sets.

3 Previous Work on Automatic Dialogue Act Recognition

The DA recognition task comprises two related sub-tasks: segmentation, and classification or tagging. These tasks may be performed jointly or sequentially. In a sequential approach the conversation is first segmented into unlabelled DA segments, then each detected segment is tagged with a DA label. The joint approach performs both tasks concurrently, detecting DA segment boundaries and assigning labels in a single step. The joint approach is able to examine multiple segmentation and classification hypotheses in parallel, whereas only the most likely segmentation is supplied to the DA classifier in a sequential approach. The joint approach is potentially capable of greater accuracy, since it is able to explore a wider search space, but the optimization problem can be more challenging. In a sequential system the two sub-tasks can be optimised independently. Note that an integrated system may be used as a segmenter by ignoring its classifications. For purposes of comparison, often it may also be used as a classifier, by forcing a human DA segmentation onto it.

Most previous work concerned with DA modelling has focused on tagging presegmented DAs, rather than the overall recognition task which includes segmentation and tagging. Indeed, automatic linguistic segmentation [Stolcke and Shriberg, 1996, Shriberg et al., 2000, Baron et al., 2002] is often regarded as a research problem itself.

3.1 Automatic Dialogue Act Tagging

The use of a generative HMM discourse model [Nagata and Morimoto, 1993], in which observable feature streams are generated by hidden state DA sequences, has underpinned most approaches to DA modelling, and a good overview of this approach is given by Stolcke et al. [2000]. The discourse history is typically modelled using an n-gram over DAs, although approaches such as polygrams [Warnke et al., 1997] have been tested. Lexical features have been widely used for DA tagging (section 3.3), via cue words or statistical language models, including approaches such as multiple parallel n-grams [Venkataraman et al., 2005], hidden event language models [Zimmermann et al., 2006a], and factored language models [Ji and Bilmes, 2005]. Several authors have previously investigated the use of prosody to disambiguate between different DAs with a similar lexical realisation [Bhagat et al., 2003], and investigated approaches to automatically select the most informative features [Shriberg et al., 1998, Hastie et al., 2002]. Prosodic features such as duration, pitch, energy, rate of speech and pauses have been successfully integrated into the processing framework.

Ji and Bilmes [2005] have proposed a switching-DBN based implementation of the HMM approach above outlined, applying it to the DA tagging task on ICSI meeting data. They also investigated a conditional model, in which the words of the current sentence generate the current dialog act (instead of having dialogue acts which generate sequence of words). DA tagging experiments have been performed both using multiple parallel n-grams or adopting a FLM with two factors: words and DA labels. The generative approach prevails over the conditional model, reporting the best classification accuracy when used in conjunction with a FLM. Since this work used only lexical features, and a large number of DA categories (62), a direct comparison with the results provided by [Ang et al., 2005,

Zimmermann et al., 2006a, Dielmann and Renals, 2007a, Zimmermann et al., 2006b] is not possible.

Venkataraman et al. [2003] proposed an approach to bootstrap a HMM-based dialogue act tagger from a small amount of labeled data followed by an iterative retraining on unlabeled data. This procedure enables a tagger to be trained on an annotated corpus, then adapted using similar, but unlabeled, data. The proposed tagger makes use of the standard HMM framework, together with dialogue act specific language models (3-grams) and a decision tree based prosodic model. The authors also advance the idea of a completely unsupervised DA tagger in which DA classes are directly inferred from data.

More recently, there have been a number of conditional models applied to DA classification including support vector machines (SVMs) [Fernandez and Picard, 2002, Liu, 2006] and maximum entropy classifiers [Venkataraman et al., 2005, Ang et al., 2005]. Features for these models include both lexical and prosodic cues, as well as contextual DA information [Venkataraman et al., 2005] (table 5).

A framework for the automatic DA classification of the Spanish CallHome spontaneous speech corpus (using 8 DA labels) has been outlined by Fernandez and Picard [2002]. The proposed approach relies on a SVM based classifier and a set of features derived from energy and pitch contours. Numerical results show the importance of prosodic cues, highlighting how even without a lexical transcription it is still possible to detect DAs well above chance.

Liu [2006] proposed an automatic DA classifier based on the combination of multiple binary SVM classifiers via Error Correction Output Codes. This work extends the 5 DA NIST tagging task outlined in Ang et al. [2005] comparing the originally adopted maximum entropy classifier with a multiclass SVM and 4 different setups based on ECOC SVM classifiers. All ECOC classifiers perform better than a multiclass SVM, but unfortunately they are not able to outperform the baseline MaxEnt system of Ang et al. [2005].

Generative and conditional approaches can also be combined: for example Surendran and Levow [2006] integrated local discriminative SVM classifiers (using prosodic and lexical features) within an HMM framework by applying Viterbi decoding to class posterior probabilities estimated using the SVMs. The SVM-HMM system has been applied to the 13 DA classes Maptask corpus [Carletta et al., 1997] consisting in dialogues between two participants interacting on a game-move task: a *giver* provides instructions to guide a *follower* through the route on a map.

3.2 Automatic Dialogue Act Recognition

An early system for the integrated joint DA segmentation and classification has been outlined by Warnke et al. [1997]. 18 DA classes are automatically recognised in short task oriented two person conversations (appointment scheduling of the German VERBMOBIL corpus). The system using: a multi-layer perceptron and a Language Model for segmentation, a polygram LM for DA classification, and a joint search algorithm to score multiple joint recognition hypotheses; reports an improvement over a sequential approach.

[Ang et al., 2005] addressed the automatic dialogue act recognition problem using a sequential approach, in which DA segmentation was followed by classification of the candi-

date segments. Promising results were achieved by integrating a boundary detector based on *vocal pauses* (table 6) with a hidden-event language model HE-LM (a language model including dialogue act boundaries as pseudo-words). The dialogue act classification task was carried out using a maximum entropy classifier, together with a relevant set of textual and prosodic features. This system segmented and tagged DAs in the ICSI Meeting Corpus (using the 5 broad DA categories outlined in section 2.1), with relatively good levels of accuracy. However results comparing manual with automatic ASR transcriptions indicated that the ASR error rate resulted in a substantial reduction in accuracy.

In a later work Zimmermann et al. [2006a] compared two joint approaches on the same experimental setup. An extended HE-LM able to predict not only DA boundaries but also the type of the DA, and a HMM recogniser inspired by HMM based part of speech taggers, was trained on lexical features and compared using several of the metrics discussed in section 4. The joint HE-LM system obtained lower recognition error rates than the HMM based DA recogniser, achieving performances closer to the discriminative sequential approach of Ang et al. [2005].

A further extension of the joint HE-LM DA recogniser introduced by Zimmermann et al. [2006a] has been developed in Zimmermann et al. [2006b]. A discriminative maximum entropy DA boundary detector and tagger is trained on discretised inter-word pauses with a lexical context of 4 words. Then the weighed combination of the classification probabilities for both systems (HE-LM and MaxEnt) provides the most likely sequence of labelled DA units. Experimental results on the ICSI 5 DA tasks suggest that the novel combined approach is capable of better recognition performances than the sequential approach of Ang et al. [2005]. Note that multiple concurrent DA segmentation and classification hypotheses could be evaluated by joint DA recognisers, enabling the investigation of larger search spaces compared with two-step sequential segmentation-classification approaches.

An integrated framework for the joint DA segmentation and tagging has been outlined by Dielmann and Renals [2007a]. The proposed system is based on: a switching dynamic Bayesian network (DBN) architecture, a set of features related to lexical content and prosody, and a Factored Language Model. The switching DBN coordinates the recognition process by integrating all the available resources. Experiments on the 5 broad DA categories of the ICSI meeting corpus have been carried out, using both manually transcribed speech, and the output of an automatic speech recogniser, and using different feature configurations. The DA segmentation and recognition results are similar to those of Ang et al., although using a discriminative MaxEnt DA classifier [Ang et al., 2005] resulted in a 5% lower error rate for the tagging task. Experiments on the AMI corpus using an extended version of the switching DBN framework have been reported in Dielmann and Renals [2007b].

3.3 Features for Automatic Dialogue Act Processing

Table 5 lists some of the features used in previous works to perform automatic DA classification; while table 6 shows the most frequently used features that have been adopted for the DA segmentation task.

The most common features used for the automatic DA segmentation and classification can be subdivided in:

Feature / Article	Ang et al. [2005]	Rosset and Lamel [2004]	Fernandez and Picard [2002]	Rotaru [2002]	Lendvai et al. [2003]	Andernach [1996]	Reithinger and Klesen [1997]	Venkataraman et al. [2002]	Venkataraman et al. [2003]	Keizer and Akker [2005]	Venkataraman et al. [2005]	Jurafsky et al. [1998]	Zimmermann et al. [2006a]	Zimmermann et al. [2005]	Warnke et al. [1997]	Katrenko [2004]	Webb et al. [2005]	Ji and Bilmes [2005]	Surendran and Levow [2006]	Liu [2006]	Dielmann and Renals [2007b]	Verbree et al. [2006]
Sentence length	✓									✓	✓									✓	✓	✓
First two words	✓	✓									✓									✓		
Last two words	✓										✓									✓		
Number of utterances		✓																				
Bigrams of words in segment				✓																		
Bigram of first two words																				✓		
Utterance type						✓																
Presence/absence Wh-words						✓																
Subject Type						✓																
Specific cue words/phrases						✓						✓				✓						✓
First verb type						✓																
Second verb type						✓																
Question mark						✓																✓
Sparse bag of ngrams																			✓			
Specific patterns										✓												
Grammar pattern										✓		✓										
Polygrams of words							✓								✓							
Factored Language Model																		✓			✓	
Part Of Speech ngrams																						✓
Ngrams of words								✓	✓		✓	✓	✓	✓			✓	✓		✓		✓
First word of next segment	✓										✓									✓		
Speaker (turn) change		✓							✓		✓								✓	✓		
Words in last 10 DA's					✓							✓			✓							
Pitch			✓		✓				✓			✓			✓				✓		✓	
Energy			✓		✓							✓							✓		✓	
Duration			✓		✓				✓			✓			✓				✓		✓	
Pauses					✓				✓			✓			✓						✓	
Rate of speech					✓														✓			
Ngrams of previous DA's								✓	✓		✓			✓				✓	✓		✓	✓
Previous DA hyp. / posteriors		✓								✓												
Next DA										✓												
Previous 10 DAs (from ref.)					✓																	

Table 5: Features used for automatic DA-classification in different studies

Feature / Article	Kolar et al. [2006]	Stolcke and Shriberg [1996]	Lendvai and Geertzen [2007]	Zimmermann et al. [2006b]	Dielmann and Renals [2007b]	Ang et al. [2005]	Zimmermann et al. [2006a]
Segmentation only	✓	✓					
Surrounding Words				✓			
Ngrams of words	✓	✓				✓	✓
Part Of Speech ngrams		✓					
Tokenized Words			✓				
Bag of Words			✓				
Word relevance					✓		
Factored Language Model					✓		
Disfluencies			✓				
Repeats	✓						
Overlapping Speech			✓				
Pauses	✓		✓	✓	✓	✓	
Pitch	✓				✓		
Duration	✓				✓		
Energy	✓				✓		

Table 6: Features used for automatic DA-segmentation in different studies.

Lexical features usually a language model based on words: DA specific ngrams of words, polygrams, factored language models, part-of-speech ngrams, etc. Some systems also rely on selected cue words/phrases and specific lexical or grammatical patterns. The number of words contained by the current DA segment (sentence length) is also a lexical related feature frequently adopted for DA classification. In order to evaluate fully automatic DA tagging and recognition systems, automatic ASR transcriptions are required. Inaccuracies of the automatically recognised speech have an adverse effect on lexical derived features. Therefore it is worth evaluating the full system both on manual and automatic transcriptions in order to estimate the overall degradation of performances caused by the ASR output.

Context features describe the relation between the current and the surrounding utterances, e.g. to indicate temporal overlap between speakers.

Prosodic features represent a wide group of acoustic related features like: F0 and pitch slopes, the duration of words, unvoiced pauses, speech rate, features derived from spectral coefficients, etc.

A discourse model (or discourse grammar) is based on the DA types of the preceding or surrounding segments. It is important to note whether this history is maintained on

the actual output of the DA classifier, or on the hand-annotated DAs. For a realistic evaluation, the actual classification results should be used; however, generating the history from annotated DAs gives an estimation of the potential usefulness of this kind of features.

Two important aspects related to the feature extraction process are source and scope of the extracted features. Even if all the information required for feature extraction should come from fully automatic approaches, several systems are trained on features relying on manually labelled data. Moreover many systems are frequently evaluated using features based on manual annotations (i.e: lexical features estimated using the reference orthographic transcriptions), either because data from an automatic system are not available yet, or to assess the potential usefulness of a new feature family. Automatic DA processing is often a component block of a larger infrastructure (section 5), therefore specific constraints imposed by the applicative domain have a deep influence on the feature scope. For example, in a meeting browsing application designed to offer its facilities online during an undergoing meeting, the DA recognition process will have access only to the past conversations. Note also that in this application the DA processing should operate in real-time relying on a less accurate ASR transcription. In a post-processing application (e.g., offline meeting corpus browser), the whole discourse is available, allowing the use of features which look ahead in the time.

4 Metrics and Evaluation

Each of the segmentation, classification and the joint segmentation and classification tasks, has its own set of performance metrics. If performance evaluation is straightforward for the DA tagging task, the same cannot be said about DA segmentation or recognition tasks. Several evaluation metrics have been proposed, but the debate on this topic is still open. Moving from the NIST-SU error metric introduced in NIST website [2003], several DA segmentation and recognition metrics have been proposed by Ang et al. [2005] and subsequently extended by Zimmermann et al. [2006a].

4.1 Classification metrics

The performance of DA classification using manually annotated segments is usually measured in terms of accuracy, which is the percentage of correctly classified segments, or classification error rate, which is the percentage of incorrect classifications. For a more detailed evaluation, occurrences and correct classifications of each DA class can be counted separately [Lesch et al., 2005a]:

$$
\begin{aligned}
correct_{DA} &= \textit{the number of times DA was correctly classified} \\
annotated_{DA} &= \textit{the number of occurrences of DA in the annotated test data} \\
tagged_{DA} &= \textit{the number of times DA was classified}
\end{aligned}
$$

Based on these counts, we define the recall ($Recall_{DA}$) and precision ($Precision_{DA}$) measures for each DA class, as well as the accuracy and mean precision for the whole test

set:

$$
\begin{aligned}
Recall_{DA} &= \frac{correct_{DA}}{annotated_{DA}} \\
Precision_{DA} &= \frac{correct_{DA}}{tagged_{DA}} \\
Accuracy &= \frac{\sum_{DA} correct_{DA}}{\sum_{DA} annotated_{DA}} \\
Precision &= \frac{\sum_{DA} Precision_{DA} * annotated_{DA}}{\sum_{DA} annotated_{DA}}
\end{aligned}
$$

4.2 Segmentation metrics

The evaluation of the automatic DA segmentation is a non-trivial task. Several evaluation metrics can be defined, each giving a different perspective on the segmentation results. Figure 1 illustrates the principal metrics used to evaluate the accuracy of automatic DA segmentation. NIST-SU, recall, precision, f-measure and boundary are based on boundaries. Each word is followed by a potential boundary position, and segmentation is a binary classification into boundaries and non-boundaries. There are four possible outcomes: boundaries may be correctly identified (true positives, tp) or missed (false negatives, fn), non-boundary positions may be correctly identified (true negatives, tn) or a false boundary may be hypothesised (false positives, fp). The sum $tp + tn + fp + fn$ is equal to the number of words. The occurrences of these four events are counted. The boundary-based metrics take different combinations of these counts into consideration:

$$
\begin{aligned}
NIST - SU &= \frac{fp + fn}{tp + fn} \\
Boundary &= \frac{fp + fn}{tp + tn + fp + fn} \\
Recall &= \frac{tp}{tp + fn} \\
Precision &= \frac{tp}{tp + fp}
\end{aligned}
$$

The F-measure is the harmonic mean of the computed precision and recall given the reference sentence boundaries and the boundaries hypothesised by the segmentation system: $F = 2 \times Recall \times Precision/(Recall + Precision)$. The other two segmentation metrics, DA segment error rate (DSER) and Strict, are based on segments. DSER is the fraction of reference segments which have not been correctly recognised, meaning that either of the boundaries is incorrect. Strict is a variant of DSER in which each DA segment is weighted with its length (number of words).

4.3 Joint segmentation and classification metrics

The DA recognition task is more challenging, since the limited accuracy of automatic segmentation and classification are combined together. Note that a direct comparison between DA recognition and classification results is difficult. However the DA classification

Reference	S|Q.Q.Q.Q|S.S.S|B|S.S|
System	S|Q|S|Q.Q|D.D.D|S.S|S|
NIST-SU	.c.e.e...c.....c.e.e.c
Boundary	.c.e.e.c.c.c.c.c.e.e.c
Recall	.c.......c.....c.e...c
Precision	.c.e.e...c.....c...e.c
DSER	c|...e...|..c..|e|.e.|
Strict	c|e.e.e.e|c.c.c|e|e.e|

Metric	Counts	Reference	Rate
NIST-SU	3 FP, 1 miss	5 boundaries	80%
Boundary	3 FP, 1 miss	11 (non-)boundaries	27%
Recall	4 correct	5 boundaries	80%
Precision	4 correct	7 hypothesised boundaries	57%
F-Measure	-	-	67%
DSER	3 match errors	5 reference DAs	60%
Strict	7 match errors	11 reference words	63%

Figure 1: Metrics for segmentation based on boundaries (NIST-SU, Recall, Precision, F-Measure and Boundary) and on segments (DSER and Strict). The symbol '|' is used to indicate boundaries between consecutive DAs and '.' stands for non-boundaries between words. The letters S, Q, D, and B represent single words of the DAs. Correctly hypothesised boundaries are marked with a letter c while e is used to label false positives and missed boundaries.

performance can be interpreted as an upper boundary for the whole recognition process, which would be reached if automatic segmentation was perfect.

A set of metrics, in analogy to the segmentation metrics of section 4.2, can be defined for the recognition task. Figure 2 illustrates a set of performance metrics for joint segmentation and classification of DAs. In contrast to the NIST error metric for segmentation, the hypothesised DA label is taken into account as well, leading not only to false positives (insertions) and misses (deletions) but also to substitutions. While the strict error metric requires correct DA boundaries the lenient metric completely ignores segmentation errors. As the DER can also be defined via a DA based recall, DA based precision can be defined as well, leading to a DA based F-measure: $F = 2 \times Recall \times Precision / (Recall + Precision)$. Note that recall, precision and F-measure are based on dialogue act units, not on DA boundaries as it was for the segmentation metrics.

While higher values for Recall, Precision and the F-measure indicate higher performances, the remaining metrics are error metrics, thus higher values imply lower performances. It is important to note that these metrics and all evaluations presented in this chapter are intrinsic, being purely based on the comparison between human annotation and classifier/recogniser output. Knowledge of the discourse structure could be beneficial in several applicative domains (section 5); thus the automatically classified/recognised DAs often form the input of further processing stages. However the effects of DA segmentation errors and DA misclassifications on the overall system performances depend on how the DA recogniser output was used. These effects are not taken into account by the metrics defined

Reference	S|Q.Q.Q.Q|S.S.S|B|S.S|
System	S|Q|S|Q.Q|D.D.D|S.S|S|
NIST	.c.e.e...c.....e.e.e.c
Strict	c.e.e.e.e.e.e.e.e.e.e.
Lenient	c.c.e.c.c.e.e.e.e.c.c.
DER/Recall	c|...e...|..e..|e|.e.|
Precision	c|e|e|.e.|..e..|.e.|e|

Metric	Counts	Reference	Rate
NIST	3 FP, 1 miss, 1 subst.	5 boundaries	100%
Strict	10 words	11 words	91%
Lenient	5 words	11 words	45%
DER	4 erroneous dialog acts	5 dialog acts	80%
Recall	1 correct dialog act	5 dialog acts	20%
Precision	1 correct dialog act	7 dialog acts	14%
F-Measure	-	-	17%

Figure 2: Metrics for joint segmentation and classification: the boundary based NIST error rate, the word based strict and lenient metrics, as well as the DA error rate (DER). The DA based recall, precision, and corresponding F-measure are illustrated in the lower part of table. The symbol '|' is used to indicate boundaries between consecutive DAs and '.' stands for non-boundaries between words. The letters S, Q, D, and B represent single words of the same DA unit; S, Q, D, and B also represent the dictionary of 4 possible DA labels. Correctly recognised elements are marked with a letter c while e is used to mark errors.

in table 1 and 2, and are not examined here. Ideally, the users of a DA segmenter/classifier should separately investigate the effects of different DA recognition errors. Given such analysis, the most appropriate metric can be identified, and the DA recognition system can be optimised for this specific application.

4.4 Evaluation on Automatic Speech Recogniser output

The reference DA annotation is produced on top of the manually transcribed word sequence. When the reference orthographic transcription is replaced by the ASR output, the DA tags need to be applied to a different word sequence, owing to ASR errors. Since a manual re-annotation of the ASR output would be extremely expensive, the evaluation scheme proposed by Ang et al. [2005] is often adopted: ASR words are mapped into the manually annotated segments according to their midpoint $0.5 * (word\_start\_time + word\_end\_time)$, thus inheriting their reference DA labels.

Insertions and deletions Since the proposed alignment method is segment-based, insertions and deletions of single words are ignored. However, insertions and deletions of entire DA segments occur if the recogniser finds words outside of the boundaries of any annotated dialogue act, or if no words are recognised within the boundaries of an annotated DA. For example in the AMI meeting corpus, automatic transcriptions are available for 101585 dialogue acts; the *midpoint alignment* results in 91537 annotated dialogue

act segments with recognised words, and 9968 empty DA segments without words. Although this is a large fraction, the information loss is likely to be less severe, as 66% of the deleted segments contain only laughs, coughs and other non-speech noises; 70% are of type Fragment and have no function in the discourse. While 49.2% of the segments of type Fragment are deleted, the loss on all other types is less severe, between 1% and 7%. Only 14% of the deleted segments are non-Fragments containing more than one word.

The deleted DA segments can be considered in three different ways:

Include deletions as misclassifications Often in the ASR output there is no indication that a dialogue act has taken place unless words from it were recognised. Therefore deleted segments will be scored as errors.

Classify deletions However through automatic Speaker Activity Detection it is possible to estimate if a participant spoke, even when no words were recognised by the ASR system. Therefore it is possible to include these segments as ordinary dialogue acts without words. They can be classified using non-lexical features like the duration, overlap with previous DAs, or prosody related features. Classifiers which are limited to lexical features can choose the most frequent class (or the most frequently lost class, e.g. Fragment). Note that this type of evaluation allows a closer comparison to results on manual transcriptions.

Exclude deletions Deletions can be excluded from the accuracy metrics, showing the potential of the DA classifier on ASR words more clearly.

Impact of ASR on DA classification DA tagging experiments both on ICSI [Ang et al., 2005, Dielmann and Renals, 2007a] and AMI data show that the classification accuracy on automatically recognised words is approximately 7–10% (absolute) lower than on reference transcriptions.

5 Applications of Automatic Dialogue Act Processing

Dialogue acts form a useful level of representation for the interpretation of conversations, providing a bridge between an orthographic (word-level) transcription, and a richer representation of the discourse. DA labels may incorporate syntactic, semantic and pragmatic factors: in addition to providing information about the structure of a dialogue and the course of a conversation, DAs are also able to capture, at a coarse level, individual speaker attitudes and intentions, their interaction role and their level of involvement. The reliable recognition of the DA sequence in a conversation, and the resulting knowledge of the discourse structure, can be beneficial in the development of applications in a multitude of domains, such as: spoken dialogue systems, machine translation, automatic speech recognition, automatic summarisation, topic segmentation and labelling, action items detection, group action detection, participant influence detection, and dialogue structure annotation.

As outlined in section 2, during the last decade, multiple corpora have been annotated in terms of DAs, and a relevant literature about automatic DA recognition (section 3) has been developed. Several works also focused on the exploitation of the automatically extracted DAs. Moving from the idea that the knowledge about the ongoing conversation

(conveyed by DAs) can be used to enhance language modelling; improving Automatic Speech Recognition of conversational speech was one of the first targets. Jurafsky et al. [1997a] investigated the use of automatically detected Dialogue Acts to improve Automatic Speech Recognition. The 1155 pre-segmented conversations from the Switchboard database were automatically tagged using the clustered dictionary of 42 DA labels. The system made use of a generative DA tagging infrastructure based on: prosodic features (pitch, speaking rate, energy, etc.), 42 word sequence based trigram models, and a bigram discourse language model. Automatic transcriptions were generated through ASR and then fed to the automatic DA tagger. The automatically detected DA classes are then used to rescore the ASR output by means of a novel *DA conditioned mixture Language Model*: N-best lists associated to each test-set utterance have been rescored using a mixture of DA specific LMs. Numerical results on the Switchboard corpus show only a limited improvement (0.3%) on the ASR word error rate because of the skewed distribution of DA classes (statements account for 83% of the corpus). However DA rescored ASR should have a larger impact on specific tasks with more even DA distributions (e.g., task oriented dialogs). A deeper analysis and further generalisations (*mixture of posteriors*) of the *mixture of language models* have been reported in Stolcke et al. [2000]. Related experiments on Maptask [Taylor et al., 1998], show that the automatic choice of the most appropriate language models from a set of 12 DA specific LMs (detected using intonation modeling), can improve the speech recognition word error rate by an absolute 1%.

Machine translation is another applicative domain where DA recognition can be invaluable, since DAs can help resolve ambiguities in translating utterances. The VerbMobil project investigated machine translation in dialogue systems [Küssner, 1997, Wahlster, 2000], similarly to the work independently done by Lee et al. [1997]. The use of DAs for machine translation of spoken task-oriented dialogues has been also proposed in the context of the C-STAR project by Levin et al. [2003].

Automatic detection of *action items*, intended as public commitments to perform a defined task, is a novel research topic which share some analogies with and relies on automatic DA recognition. In the work of Purver et al. [2007], 4 task specific Action Item Dialogue Acts (description, time-frame, owner and agreement) are automatically detected combining 4 independent SVM classifiers trained on: lexical, prosodic features and conventional ICSI DA tags. The automatically detected AIDAs are then rule-based parsed and summarised in order to outline the identified action items. Disambiguating the pronoun *you*, between its generic and referential use in a conversation, is a task related to *action items* detection, which could be useful to identify the owner of an action item (who committed to perform a given task). The SVM based system proposed by Gupta et al. [2007b], based on DAs, lexical and part of speech features, is able to disambiguate the two uses with an accuracy of 84.4% on 2 person conversations from the Switchboard corpus. This represents a significant result, well above the baseline 56.4% achievable always predicting the dominant class. In particular DAs proved to be crucial for this task, reaching an accuracy of 80.92% even if used alone. Later experiments [Gupta et al., 2007a], using a similar setup on the AMI corpus, reported an accuracy of 75.1% with the full feature setup and 71.9% using only DAs (dominant class baseline of 57.9%).

Automatically detecting when decisions are reached during a conversation is another target application for automatic DA extraction. Hsueh and Moore [2007] used both DA unit

temporal boundaries and DA labels for automatic decision detection in conversational speech. The manually annotated DA units are classified as decision making DAs or non-decision DAs using a MaxEnt classifier and a rich set of lexical, prosodic, topical and contextual features (like speaker role and DA labels). Experiments on the AMI corpus show that decision making DAs can be detected with a precision of about 72% (66% using only contextual features).

Differently from written text, automatically transcribed speech lacks of a proper punctuation. It is often unpractical to process the entire raw transcription or to evaluate the resulting system on unsegmented data, thus shorter speech segments need to be defined. The temporal boundaries of automatically recognised DA units provide a principled way to segment conversational speech. For example Murray et al. [2006] and Murray and Renals [2006] adopted the DA segments as the atomic unit for automatic extractive summarisation; features like lexical cues, speaker activities and term frequencies were individually extracted from each DA unit, and Singular Value Decomposition carried out on the resulting DA based feature vectors. Note that although DA segments are a good solution for automatic speech segmentation, some low-level segmentation techniques such as "Spurts" [Baron et al., 2002], continuous speech segments separated by at least half a second of silence, could represent a viable option.

Complex integrated applications based on automatic DA processing are being currently investigated. For example, topic segmentation and extractive summarisation have been combined in the "AMI Meeting Facilitator" system [Murray et al., 2007], a visual application focused on supporting offline meeting browsing. Here dialogue acts, being exploited by both subtasks (segmentation and summarisation), offer a common ground for the whole system.

6 DA Tagging, Segmentation and Recognition of the AMI Meeting Corpus

The DA tagging and recognition experiments conducted on the AMI meeting corpus extend and adapt the previous experiences acquired on former multiparty conversational corpora, like the ICSI meeting corpus. In order to compare DA classification performances on different meeting data (Switchboard, ICSI and AMI) a portable DA tagger has been developed by Verbree et al. [2006]. The proposed system makes use of several feature families: question marks and lexical cues, unit lengths, compressed ngrams of both words and POS tags; and a bigram discourse model. The extracted features are then modelled using the J48 classifier of the Weka toolkit. While the classification accuracy achieved on the Switchboard 42 DA task is about 5% lower than the state of the art, the system outperforms all the previous works on the 5 DA ICSI task, reaching an accuracy of 89.3%. The classification accuracy on the AMI 15 DA is about 59.8% using reference orthographic transcriptions and 49.3% using the ASR output.

The maximum entropy (MaxEnt) based classification system outlined in [Lesch, 2005, Lesch et al., 2005b] adopts a wide set of features belonging to the following 5 classes: lexical features, DA unit length and duration, temporal relation between adjacent utterances, speaker change and dialogue act history. A feature selection algorithm, which grows the

	Metric	Reference	ASR output
S	NIST-SU	20.4	26.5
E	DSER	12.8	17.0
G	Strict	28.5	29.4
M.	Boundary	3.1	4.4
R	NIST-SU	71.3	85.9
E	DER	51.9	62.5
C.	Strict	62.1	68.5
	Lenient	42.2	48.3

Table 7: DA segmentation and recognition error rates (%) on the AMI meeting corpus both on reference manual transcriptions and ASR output; segmentation results are reported using the interpolated FLM, whenever the hybrid FLM+iFLM system has been used for the joint DA recognition task.

feature subset by iteratively ranking the features according to their classification accuracy, has been adopted to select only the most relevant features and reduce the feature set. The best classification accuracy obtained on the AMI evaluation set is 65.8% for reference words and 54.9% with automatically recognised words (classifying ASR deleted DA units by chance). This result defines the state of the art for the 15 DA AMI tagging task. Similarly to Verbree et al. [2006] and Dielmann and Renals [2007b], when the reference transcription is replaced by the ASR output, the classification accuracy falls by about 10% (absolute).

The discriminative MaxEnt approach outperforms the generative FLM based classifier of Dielmann and Renals [2007b] by about 6% both on reference (59.1%) and automatic transcriptions (49.3%). However the switching DBN infrastructure outlined in Dielmann and Renals [2007b], being able to perform concurrently both DA segmentation and classification, is principally targeted to the joint DA recognition task rather than being forced to classify presegmented data. DA recognition experiments have been reported using three different language model configurations: an FLM trained only on AMI data, a weighted interpolated FLMs trained also on ICSI and Fisher data, and an hybrid setup with both an FLM and an interpolated FLM. The interpolated FLM, thanks to its richer dictionary and language model, reduces the number of segmentation errors by a factor of 2–3, at the cost of a slightly degraded DA classification accuracy. A hybrid approach, using both FLMs, allows a trade off between segmentation and classification, improving the overall recognition accuracy. Note also that joint DA recognition approaches perform segmentation and classification in a single and indivisible process, such that adjustments which improve the segmentation may lead to lower classification accuracy and vice-versa. The reported experiments (table 7) suggest that it is possible to perform automatic segmentation into DA units with a relatively low error rate. However the operation of automatic recognition into 15 imbalanced DA categories has a relatively high error rate, indicating that this remains a challenging task.

Both DA tagging and automatic DA recognition are open research topics, thus further investigations and improvements both on the feature extraction process and on the statistical modelling framework will be discussed in the next paragraphs.

Features Language models automatically derived from text corpora are typically very large, with up to several hundred thousand n-gram features. The system outlined by Verbree et al. [2006] presents an approach to select a small number of lexical cues, which shows relatively good classification accuracies even using very small models. Shrinking the feature set helps reducing the computing overhead when a DA recogniser will be employed as part of an online application, deemed to run in real-time. In many cases a smaller model with 1-2% lower accuracy, which fits easily on a machine together with various other modules, will be preferable to a slightly better model with vast memory requirements.

Likewise, the selection of feature types investigated using a maximum entropy modelling approach [Lesch, 2005, Lesch et al., 2005b], reduces the number of binary features and provides invaluable insights to the usefulness and the overlap between different types of features. Lexical features, utterance length and duration, as well as context-dependent features positively contribute to the final results. Being the lexical features, especially word identities, utterance-initial and utterance-final words, the most salient ones.

Future research should include an even deeper analysis of the individual contributions granted by the current features; and should examine the potential of introducing new feature families:

Multi-modal features Other modalities like gestures or focus of attention may provide additional valuable clues.

Forward-looking features The experiments conducted so far on the AMI meetings make use only of backward-looking features, i.e. features derived from material up to the current utterance. Assuming that the entire meeting is available and the DA recognition framework will be part of an application which allows offline processing, nothing prevents from exploiting forward-looking features such as: word identities from the following utterance, future speaker changes, etc.

Additional information from speech activity detection Some utterances are lost in the ASR recognition process, since none of their words were recognised. However, some of these utterances can probably be recovered using automatic speaker activity detection (section 4.4).

Advanced classification methods The DA classification methods applied to the AMI task are based on "flat" models which discriminate between all 15 DA types in one step. However it is possible to combine multiple specialised classifiers creating a hierarchically layered classifier.

Models with fewer classes are often more discriminative than models with a large number of classes. One way to take advantage of this is to group the classes and perform the final classification in two or more steps. DA types which are similar, or frequently confused, can be merged into one abstract class, resulting in a model with fewer classes. When this model predicts an abstract class, a secondary model can be used, which is trained to discriminate between the subclasses which were collapsed into the abstract class.

Another approach targeted on reducing the number of classes is based on the notion of dialogue act dimensions: each DA type can be described by a tuple of dimensions, each of which has a small set of values. Each dimension can be modeled separately, and a

meta model can be applied to map a tuple of values to an actual DA type. While the dialogue act labels of the ICSI-MRDA scheme are clearly composed of one or more tags which represent orthogonal properties of the utterance, the AMI DA scheme consists only of a flat list of 15 DA types. However, it is still possible to identify various aspects, or "dimensions", in which any two of the 15 types are similar or different. For instance Elicit-Inform, Elicit-Offer-Or-Suggestion, etc. have in common that they elicit information from the other participants. On the other hand, Offer and Elicit-Offer-Or-Suggestion are similar in that both of them are concerned with offers. Thus we can hypothesise that two aspects of the AMI dialogue acts are: "information type" (inform, suggest/offer, assess, ...) and "direction" (whether the speaker expresses information, or elicits information).

References

T. Andernach. A machine learning approach to the classification of dialogue utterances. *Computing Research Repository*, July, 1996.

A. Anderson, M. Bader, E. Bard, E. Boyle, G. M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert. The HCRC map task corpus. *Language and Speech*, 34:351–366, 1991.

J. Ang, Y. Liu, and E. Shriberg. Automatic dialog act segmentation and classification in multiparty meetings. In *Proc. ICASSP*, volume 1, pages 1061–1064, Philadelphia, USA, 2005.

J. L. Austin. *How to do Things with Words.* Oxford: Clarendon Press, 1962.

E.G. Bard, C. Sotillo, A.H. Anderson, H.S. Thompson, and M.M. Taylor. The DCIEM map task corpus: Spontaneous dialogue under sleep deprivation and drug treatment. *Speech Communication*, 20(1):71–84, 1996.

D. Baron, E. Shriberg, and A. Stolcke. Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues. In *ICSLP*, Denver, Colorado, USA, September 2002.

S. Bhagat, H. Carvey, and E. Shriberg. Automatically generated prosodic cues to lexically ambiguous dialog acts in multiparty meetings. In *Proc. International Congress of Phonetic Sciences*, pages 2961–2964, August 2003.

J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation Journal*, In Press.

J. Carletta, A. Isard, S. Isard, J. C. Kowtko, G. Doherty-Sneddon, and A. H. Anderson. The reliability of a dialog structure coding scheme. *Computational Linguistics*, 23: 13–31, March 1997.

J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner. The AMI meetings corpus. In *Proceedings of the Measuring Behavior 2005 symposium on "Annotating and measuring Meeting Behavior"*, 2005. AMI-108.

A. Dielmann and S. Renals. Multistream recognition of dialogue acts in meetings. In *Proc. Multimodal Interaction and Related Machine Learning Algorithms Workshop (MLMI–06)*, pages 178–189. Springer, 2007a.

A. Dielmann and S. Renals. DBN based joint dialogue act recognition of multiparty meetings. In *Proc. IEEE ICASSP*, volume 4, pages 133–136, April 2007b.

R. Fernandez and R.W. Picard. Dialog act classification from prosodic features using support vector machines. In *Proceedings of speech prosody 2002*, April 2002.

J Godfrey, E. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520, San Francisco, March 1992.

S. Gupta, J. Niekrasz, M. Purver, and D. Jurafsky. Resolving "you" in multi-party dialog. In *SIGdial*, September 2007a.

S. Gupta, M. Purver, and D. Jurafsky. Disambiguating between generic and referential "you" in dialog. In *ACL*, June 2007b.

H. Hastie, M. Poesio, and S. Isard. Automatically predicting dialogue structure using prosodic features. *Speech Communication*, (36):63–79, 2002.

Y. Horiuchi, Y. Nakano, H. Koiso, M. Ishizaki, H. Suzuki, M. Okada, M. Naka, S. Tutiya, and A. Ichikawa. The design and statistical characterization of the japanese map task dialogue corpus. *Journal of Japanese Society for Artificial Intelligence*, 14(2):261–272, 1999.

P. Hsueh and J. Moore. What decisions have you made: Automatic decision detection in conversational speech. In *NACCL/HLT*, pages 25–32, Rochester, NY, USA, April 2007.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The ICSI meeting corpus. In *Proceedings of IEEE ICASSP 2003, Hong Kong, China*, pages 364–367, April 2003.

G. Ji and J. Bilmes. Dialog act tagging using graphical models. In *Proc. ICASSP*, volume 1, pages 33–36, Philadelphia, USA, 2005.

D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. Ess-Dykema. Automatic detection of discourse structure for speech recognition and understanding. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 88–95, Santa Barbara, CA, US, 1997a. IEEE CS.

D. Jurafsky, E. Shriberg, and D. Biasca. Switchboard SWBD-DAMSL shallow-discourse-function annotation (coders manual, draft 13). Technical report, Univ. of Colorado, Inst. of Cognitive Science, 1997b. URL `http://www.icsi.berkeley.edu/cgi-bin/pubs/publication.pl?ID=001359`.

D. Jurafsky, E. Shriberg, B. Fox, and T. Curl. Lexical, prosodic, and syntactic cues for dialog acts. In Manfred Stede, Leo Wanner, and Eduard Hovy, editors, *Discourse Relations and Discourse Markers: Proceedings of the Conference*, pages 114–120. Association for Computational Linguistics, Somerset, New Jersey, 1998. URL `citeseer.ist.psu.edu/article/jurafsky98lexical.html`.

S. Katrenko. Textual data categorization: back to the phrase-based representation. In *Proceedings in 2nd International IEEE Conference "Intelligent systems", Vol. III*, pages 64–67, June 2004.

S. Keizer and R. op den Akker. Dialogue act recognition under uncertainty using bayesian networks. *Natural Language Engineering*, 1:1–30, 2005.

M. Klein, N. Ole Bernsen, S. Davies, L. Dybkjær, J. Garrido, H. Kasch, A. Mengel, V. Pirrelli, M. Poesio, S. Quazza, and C. Soria. Supported coding schemes. Technical Report MATE Deliverable D1.1, EU project LE4-8370, 1998. URL `http://mate.nis.sdu.dk/about/D1.1/`.

J. Kolar, E. Shriberg, and Y. Liu. Using prosody for automatic sentence segmentation of multi-party meetings. In *Proc. TSD 2006*, volume 9, pages 629–636, 2006.

U. Küssner. Applying dl in automatic dialogue interpreting. In *International Workshop on Description Logics*, pages 54–58, Yvette, France, 1997.

J. Lee, G. C. Kim, and J. Seo. A dialogue analysis model with statistical speech act processing for dialogue machine translation. In *Spoken Language Translations EACL97 Workshop*, pages 10–15, Budapest, Hungary, 1997.

P. Lendvai and J. Geertzen. Token-based chunking of turn-internal dialogue act sequences. In *Proceedings of the 8th SIGDIAL Workshop on Discourse and Dialogue*, pages 174–181, Antwerp, Belgium, 2007.

P. Lendvai, A. van den Bosch, and E. Krahmer. Machine learning for shallow interpreta-

tion of user utterances in spoken dialogue systems. In *Proceedings of EACL-03 Workshop on Dialogue Systems: interaction, adaptation and styles of management*, pages 69–78, 2003.

S. Lesch. Classification of Multidimensional Dialogue Acts using Maximum Entropy. Diploma thesis, Saarland University, Postfach 151150, D-66041 Saarbrücken, Germany, December 2005.

S. Lesch, T. Kleinbauer, and J. Alexandersson. A new Metric for the Evaluation of Dialog Act Classification. In *Proceedings of the Ninth Workshop On The Semantics And Pragmatics Of Dialogue (SEMDIAL 2005) – DIALOR'05*, pages 143–146, Nancy, France, June 2005a.

S. Lesch, T. Kleinbauer, and J. Alexandersson. Towards a Decent Recognition Rate for the Automatic Classification of a Multidimensional Dialogue Act Tagset. In *Proceedings of the 4^{th} IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, pages 46–53, Edinburgh, Scotland, UK, August 2005b.

L. Levin, C. Langley, A. Lavie, D. Gates, D. Wallace, and K. Peterson. Domain specific speech acts for spoken language translation. In *SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan, 2003.

Y. Liu. Using SVM and error-correcting codes for multiclass dialog act classification in meeting corpus. In *Proc. Interspeech - ICSLP*, pages 1938–1941, September 2006.

G. Murray and S. Renals. Dialogue act compression via pitch contour preservation. In *Interspeech*, Pittsburgh, USA, September 2006.

G. Murray, S. Renals, J. Carletta, and J. Moore. Incorporating speaker and discourse features into speech summarization. In *NACCL/HLT*, pages 367–374, New York, USA, June 2006.

G. Murray, P. Hsueh, S. Tucker, J. Kilgour, J. Carletta, J. Moore, and S. Renals. Automatic segmentation and summarization of meeting speech. In *NACCL/HLT*, pages 9–10, Rochester, NY, USA, April 2007.

M. Nagata and T. Morimoto. An experimental statistical dialogue model to predict the speech act type of the next utterance. *Proc. of the International Symposium on Spoken Dialogue*, pages 83–86, November 1993.

NIST website. Rt-03 fall rich transcription. http://www.nist.gov/speech/tests/rt/rt2003/fall/, 2003.

M. Purver, J. Dowding, J. Niekrasz, P. Ehlen, and S. Noorbaloochi. Detecting and summarizing action items in multi-party dialogue. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, september 2007.

N. Reithinger and M. Klesen. Dialogue act classification using language models. In *Proceedings of EuroSpeech-97*, pages 2235–2238, Rhodes, Greece, 1997.

S. Rosset and L. Lamel. Automatic detection of dialog acts based on multi-level information. In *Procedings of the ICSLP*, pages 540–543, Jeju Island, Korea, October 2004. URL `ftp://tlp.limsi.fr/public/TuB401o.2\_p540.pdf`.

M. Rotaru. Dialog act tagging using memory-based learning. Technical report, University of Pittsburgh, spring 2002. Term project in Dialogue-Systems class.

J. Searle. *Speech Acts*. Cambridge University Press, 1969.

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

E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32:127–154, September 2000.

E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, , and H. Carvey. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue, Cambridge, MA, USA*, pages 97–100, Cambridge, USA, April-May 2004.

A. Stolcke and E. Shriberg. Automatic linguistic segmentation of conversational speech. In *Proc. ICSLP*, volume 2, pages 1005–1008, October 1996.

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26:339–373, 2000. URL `citeseer.ist.psu.edu/stolcke00dialogue.html`.

D. Surendran and G. A. Levow. Dialog act tagging with support vector machines and hidden Markov models. In *Proc. Interspeech - ICSLP*, September 2006.

P. Taylor, S. King, S. Isard, and H. Wright. Intonation and dialog context as constraints for speech recognition. *Language and Speech*, 41:489–508, 1998.

A. Venkataraman, A. Stolcke, and E. Shirberg. Automatic dialog act labeling with minimal supervision. In *Proceedings of the 9th Australian International Conference on Speech Science & Technology*, December 2002.

A. Venkataraman, L. Ferrer, A. Stolcke, and E. Shriberg. Training a prosody-based dialog act tagger from unlabeled data. *Proc. of the IEEE ICASSP*, April 2003.

A. Venkataraman, Y. Liu, and E. Shriberg. Does active learning help automatic dialog act tagging in meeting data? In *Proc. Interspeech - Eurospeech*, pages 2777–2780, September 2005.

D. Verbree, R. Rienks, and D. Heylen. Dialogue-act tagging using smart feature selection; results on multiple corpora. In *IEEE Spoken Language Technology Workshop*, pages 70–73, December 2006.

W. Wahlster. *Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System*, pages 3–21. Springer, 2000.

V. Warnke, R. Kompe, H. Niemann, and E. Nöth. Integrated dialog act segmentation and classification using prosodic features and language models. In *Proc. 5th Europ. Conf. on Speech, Communication, and Technology*, pages 207–210, September 1997. Eurospeech.

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

M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke. A* based joint segmentation and classification of dialog acts in multiparty meetings. In *Proc. 9th IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 215–219, San Juan, Puerto Rico, november 2005.

M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke. Toward joint segmentation and classification of dialog acts in multiparty meetings. In *Machine Learning for Multimodal Interaction: 2nd International Workshop, MLMI 2005*, pages 187–193. LNCS 3869, Springer, 2006a.

M. Zimmermann, A. Stolcke, and E. Shriberg. Joint segmentation and classification of dialog acts in multiparty meetings. In *Proc. IEEE ICASSP*, volume 1, May 2006b.