FP6-506811

AMI

Augmented Multiparty Interaction

Integrated Project
Information Society Technologies

D5.1 Report on Initial Work in Segmentation, Structuring, Indexing and Summarization

Due date: 31/06/2005 **Submission date:** 02/09/2005
Project start date: 1/1/2004 **Duration:** 36 months
Revision: v1.0

LEAD CONTRACTOR DFKI

Project co-funded by the European Commission in the 6th Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

D5.1 Report on Initial Work in Segmentation, Structuring, Indexing and Summarization

LIST OF AUTHOR NAMES (AND AFFILIATIONS)
Tilman Becker (DFKI), editor
Jan Alexandersson (DFKI), Marc Al-Hames (TUM), Mihaela Bobeica (TNO), Jan Černocký (UBRNO), Alfred Dielmann (UEDIN), Michal Fapšo (UBRNO), Daniel Gatica-Perez (IDIAP), Yoshi Gotoh (USHFLD), Sabrina Hsueh (UEDIN), Natasa Jovanovic (UT), Thomas Kleinbauer (DFKI), Wessel Kraaij (TNO), Stephan Lesch (DFKI), Harald Lochert (DFKI), Johanna Moore (UEDIN), Gabriel Murray (UEDIN), Stephan Reiter (TUM), Steve Renals (UEDIN), Ruther Rienks (UT), Gerhard Rigoll (TUM), Petr Schwarz (UBRNO), Stanislav Sumec (UBRNO), Weiqun Xu (UEDIN), Dong Zhang (IDIAP),

Abstract: This report summarizes the work on higher level analysis in the first eighteen months of the AMI project. Analysis is done on multiple levels with an emphasis on segmentation, structuring and indexation. Based on the information extracted from meetings, we have also started work on accessing the indexed documents and generating extractive as well as abstractive summaries.

Contents

1 Introduction

One of the major goals in AMI is the deeper understanding of the structure and content of meetings and ways to make this information accessible. Building on the work in WP4 which is mainly concerned with signal-level analysis, this report summarizes the work in WP5 on higher level analysis in the first eighteen months of the AMI project. Analysis is done on multiple levels with an emphasis on segmentation, structuring and indexation. Based on the information extracted from meetings, we have also started work on accessing the indexed documents and generating extractive as well as abstractive summaries.

This deliverable attempts to summarize a multitude of work done in close cooperation across a large number of topics and project partners. Some of the work has been published already and is included here with appropriate editing to integrate it into the deliverable. The document is organized as follows:

1.1 Overview

The first sections () on segmentation and structuring detail the work done on

- dialog acts,
- addressing information,
- dominance detection,
- topic detection,
- named entity recognition,
- propositional content,
- argumentative structure,
- chunking and
- meeting group actions

On most levels, the two aspects of segmentation and classifications (or recognition) go hand in hand. For each level, we have already worked out approaches for the evaluation of our components, these mentioned in section 12. Access to information in meetings is supported in two different ways in WP5: information retrieval methods and summaries. Information retrieval is based on indexing and keyword spotting approaches which are reported in section 13. Further work on retrieval is published in [20]. Summaries are generated with two different approaches: extractive (see section 14) and abstractive (see section 15). Multimodal summaries will eventually be supported by methods for automatic video editing, see section 16.

2 Dialog Acts

The set of dialog acts used in AMI was developed in WP3. Beyond the definitions, the annotation manual[1] addresses many possible misunderstandings and provides detailed guidance for determining the correct dialog act and segment boundaries. Details on the discussion can be found on the AMI project's Wiki[2].

The following sections (2.1—2.6) are an adaptation of the paper "*Towards a Decent Recognition Rate for the Automatic Classification of a Multidimensional Dialogue Act Tagset*" by Stephan Lesch, Thomas Kleinbauer and Jan Alexandersson which was presented at the "*4th WS on Knowledge and Reasoning in Practical Dialogue Systems*" at IJCAI 2005 in Edinbourgh (see `http://www.csse.monash.edu.au/~ingrid/IJCAI05dialogueCFP.html`).

The paper presents some ideas and examinations on statistical dialogue act classification using multidimensional dialogue act labels, based on the ICSI meeting corpus and the associated MRDA tag set. Some statistics of this corpus and preliminary results of a statistical tagger for the dialogue act labels are shown. Finally, a proposal for a more realistic interpretation of these results is presented. The work is motivated by the need for a (statistical) dialogact classificator for the knowledge-based summarization performed in WP 5. Due to the initial lack of AMI data, it has been agreed to use the ICSI meeting corpus.

2.1 Introduction

A crucial capability of automatic speech processing systems is to determine the type of an utterance – question or statement or backchannel, etc. A common way to formalise this kind of information is to compile a categorisation of *dialogue acts* [3, 11] into a set of tags that meets best the requirements of the underlying task. With such a tagset it is possible to annotate a corpus of sample dialogues which can then be used as training material for a statistical classifier.

The ICSI[3] meeting recorder project [6], has developed a corpus containing roughly 72 hours of recordings of actual meetings. The corpus is fully annotated with a multidimensional tagset, which we will refer to as the MRDA tagset in this paper. A dialogue act in the MRDA set consists of a general tag, e.g. *statement* (`s`) and up to seven special tags that provide additional facets. For example, the label `qy^rt` stands for *yes-no question* with *rising tone*.

A straight-forward way to use the MRDA tagset for automatic recognition would be to treat each possible label as a monolithic unit, i.e. ignore the underlying multidimensional structure and instead understand a label merely as a string of characters. Then, after choosing a set of features and training the classifier, one can evaluate the quality of the classifier using traditional metrics like e.g. recall and precision.

Such a view, however, implies discarding useful structural information for both the classification process as well as for the evaluation. It is clear for instance that the dialogue acts `qy` and `qy^rt` are related. Therefore, if a `qy^rt`-utterance is misclassified, it makes a difference if it was classified as `qy` or as `s` - the latter did not even get the general tag correct. This effect is not reflected by traditional recall and precision measures where a classification is either correct or incorrect. Conversely, one expects an informed classifier which utilises the multidimensional properties of the MRDA tagset to yield better recognition rates than one that does not.

To verify this hypothesis, we take a closer look at the ICSI corpus. An initial investigation shows that only 82 labels occur more than 100 times and that the vast majority of the total 2050 labels occur just a few times (see figure 1). Consequently, it is very hard to use these rare acts for classification.

We have made some preliminary classification experiments and trained a maximum entropy classifier using 20000 utterances from the corpus and different variations of the tagset. This classifier was tested on a set of 14512 different utterances. We achieved 51.3% correct classifications. However, a more detailed analysis of the classification results reveals that there are another 20.2% of classifications which are assigned a less specific label,

[1] The current version can be found at `http://wiki.idiap.ch/ami/DialogueActs?action=AttachFile&do=get&target=dialogue-act-manual.07jul05.2220.pdf`

[2] `http://wiki.idiap.ch/ami/DialogueActs`

[3] International Computer Science Institute at Berkeley, CA

rank	dialogue act	count	percent
1	s	25684	23.03
2	b	14467	12.97
3	fh	6160	5.52
4	s^bk	5674	5.08
5	s^aa	4626	4.15
⋮	⋮	⋮	
29	b.%	511	0.46
30	%	460	0.41
⋮	⋮	⋮	
42	h	263	0.24
⋮	⋮	⋮	
50	h|s	193	0.17
⋮	⋮	⋮	
83	s^m	100	0.09
⋮	⋮	⋮	
1057	qy^bu^cs^d^rt	2	0.000018
1058	s^ar^bd|%	1	0.000009
⋮	⋮	⋮	
2049	qy^q^cs^d^rt	1	0.000009
2050	s:s^bk|s^rt	1	0.000009

Table 1: An excerpt from the dialogue act frequencies for the ICSI meeting corpus (Version 040317).

i.e., the correct general tag, but some special tags are missing. Additionally, 3.6% of the classifications are too specific, i.e., some special tags were assigned which are not present in the human annotation. Another 5.8% were "neighbours", which means they share a common supertype (for instance, the general tag) with the correct label.

We conclude that there is on the one hand room for improvements of the classification and the metric for evaluation could be developed to account for the "almost-hits".

In our efforts to enhance the classification results, we present an algorithm that automatically prunes the number of classes. The usage of such an algorithm is only then valid if requirements from the application at hand are incorporated, i. e., , if the application relies on the presense of a certain (additional) tag, this tag cannot be pruned.

The paper is organised as follows: the next section describes the MRDA tagset and a simplification thereof—the MALTUS tagset. In section 2.3, we discuss some of the characteristics of the ICSI meeting corpus and show how a classifier improves as the amount of training data increases. Section 2.4 details the measures used for the evaluation of classifiers and proposes a new measure. The next section describes the classification experiments. Finally, in section 2.6 we conclude the paper and provide some future directions.

2.2 Multidimensional Tagsets

The labels of a dialog act tagset are not necessarily multidimensional. The Verbmobil System, for example [1], used a small set of roughly 30 tags tailored to its particular application, the automatic translation of telephone negotiations. Examples of the Verbmobil tags are greet, bye, introduce, request, suggest.

Multidimensional tagsets, on the other hand, allow to annotate several aspects of an utterance. The DAMSL[4] tagset, for instance, defines four aspects: the communicative status, the information level and the forward and

[4]Dialogue Act Markup on Several Layers, [2]

backward looking function of the utterance. A variant of the DAMSL tagset, the SWBD tagset [5], was used for annotation in the Switchboard project; the SWBD tagset, in turn, served as the basis for the MRDA tagset [9].

The MRDA Tagset

The "Meeting Recorder Dialogue Act" tagset was used to annotate the ICSI meeting corpus.[5] Labels consist of a general tag, which may be followed by one or several special tags and a disruption mark, or of a disruption mark only. The general form is

(<general tag>(^<special tag>)?) (.<disruption mark>)?

with the following tags:

- General tags are statement (s), questions (qy/qw/qr/qrr/qo/qh), backchannel (b) and floor management (fg/fh/h).
- There are 40 special tags describing backchannels, positive, negative or uncertain responses, restatements (repetitions or corrections), politeness mechanisms and other functions.
- Disruption forms are "interrupted by other speaker" (%−) and "abandoned by speaker" (%−−). Two other tags, "indecipherable" (%) and "non-speech" (x), are included in this group.

Furthermore, there are two kinds of *compound labels*. Some utterances consist of two closely adjoining parts which constitute two DAs: e.g., a floor grabber followed by a statement can be annotated by a compound label `fg|s`. The other case is quoted speech, where labels are combined using a colon (e.g. `s:s`).

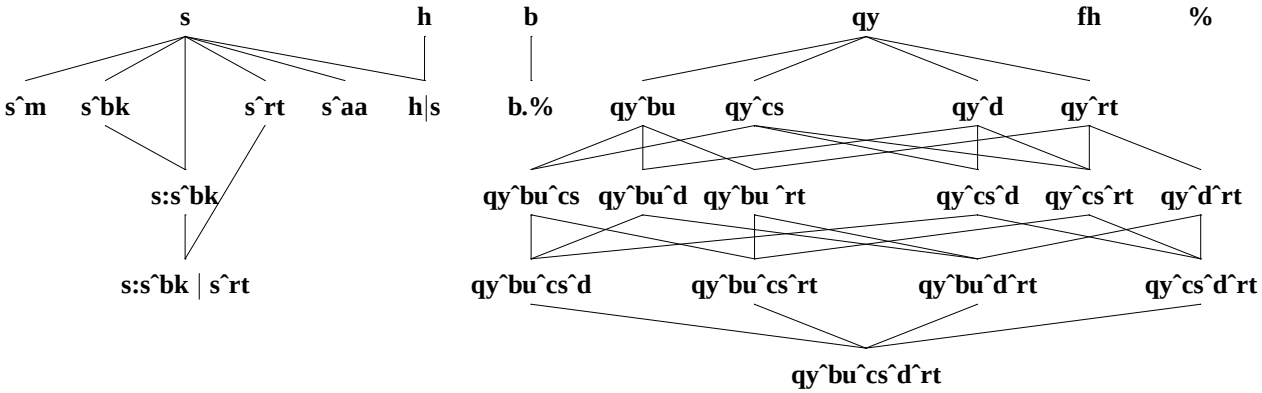

Figure 1: The lattice formed by the MRDA labels shown in table 1. Labels are ordered by the subset relation. *Compound labels*, i. e., , two labels combined with "|" or ":", are daughters of the two separate labels. Note that only the parts of the compound labels | were used in the classification experiments.

The MALTUS Tagset

MALTUS, introduced in [9], is an attempt to abstract from the MRDA tagset in order to reduce the huge number of possible labels. Several groups of MRDA tags were grouped into one MALTUS tag, and some MRDA tags were dropped altogether. An utterance is marked either as uninterpretable (U), or with one general tag (tier 1 tag, T1) and zero to five special tags (tier 2 tags, T2). Also, a disruption mark (D) may be appended. The general form of a MALTUS label is

[5] See `http://www.icsi.berkeley.edu/Speech/mr/`

(U | T1 ^ T2)? (.D)?

with the following tags:

- tier 1 tags are statement (S), questions (Q), backchannel (B) and floor holder (H).
- tier 2 tags are response types (RP/RN/RU) attention (AT), actions (DO), restated information in corrections or repetitions (RIC/RIR) and politeness (PO).

2.3 Some Corpus Characteristics

The experiments presented are based on the the ICSI meeting corpus [7], a collection of 75 meetings of roughly one hour each.

The corpus is available as text files. Each line describes one utterance: the transcribed text, the start and end times of the utterance, the time alignments of each word in the transcription, the DA label, the channel name and (optionally) adjacency pair annotation. However, the files do not contain syntactical or semantic information, POS tags or any phonological features.

The MRDA tagset theoretically allows up to several million different labels, but only some thousand of them actually occur in the corpus: the 04/03/17 version of the corpus contains 112027 utterances with 2050 different DA labels. Some of these labels are compound labels of the form a|b; we split these utterances and obtain 118694 utterances with 1256 different labels. Some utterances are explicitly marked as non-labelled (z), and some are not labelled at all; these utterances and their successors are ignored, leaving 116097 utterances from which we take the training and testing material.

Distribution of general categories over the ICSI corpus

When we map the MRDA labels to the five basic categories (statements, questions, backchannels, floor management and disruptions) in what we call "classmap 1", we see that the frequencies of these categories are very unevenly distributed - statements make up more than half of the material (See table 2). Note the descending order in the number of training examples for statements, backchannels, floor managements and questions, and how this order is reflected in the recall for these classes in a five-way classification experiment using classmap 1, see figure 4.

Category	gen. tag	%	classm.1	%
Statement	76073	64.09	66640	56.14
Backchannel	15178	12.79	14624	12.32
Floor	12276	10.34	12235	10.31
Question	8522	7.17	7374	6.21
Disruption	4113	3.47	15289	12.88
Z(nonlabeled)	2442	2.06	2442	2.06
X(nonspeech)	90	0.08	90	0.08
Σ	118694	100%	118694	100%

Table 2: Distribution of the main classes over the corpus.

Words and bigrams

We counted the number of words and bigrams over excerpts from the corpus with different sizes (with 8-fold averaging, using raw words without stemming). The logarithmic plot (see figure 2) shows that the numbers of word and bigram features keep increasing with the number of utterances examined. There is also a constant

relation between the number of words and the number of utterance-initial words—there are about five to eight times as many words as initial words. A similar relation holds between bigrams and utterance-initial bigrams.

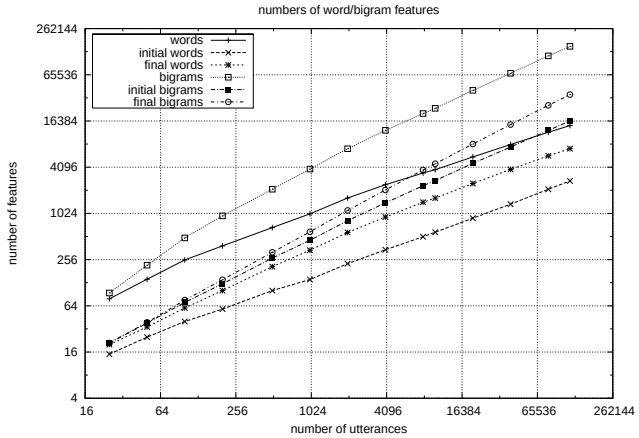

Figure 2: The number of words and bigrams for different numbers of utterances

How much training data do we need for a classifier?

With the specification of a new (MRDA-like) tagset for a corpus of meetings in mind, we were also interested in how much hand-annotated training material is needed to obtain "decent" classification using a statistical model. We found that the learning curve begins to flatten out at roughly 10000 utterances, but keeps rising with more training data. This observation (see figure 3) holds for the full set of MRDA labels, as well as when we map them to MALTUS labels, or to the five basic classes (using the "classmap 1").

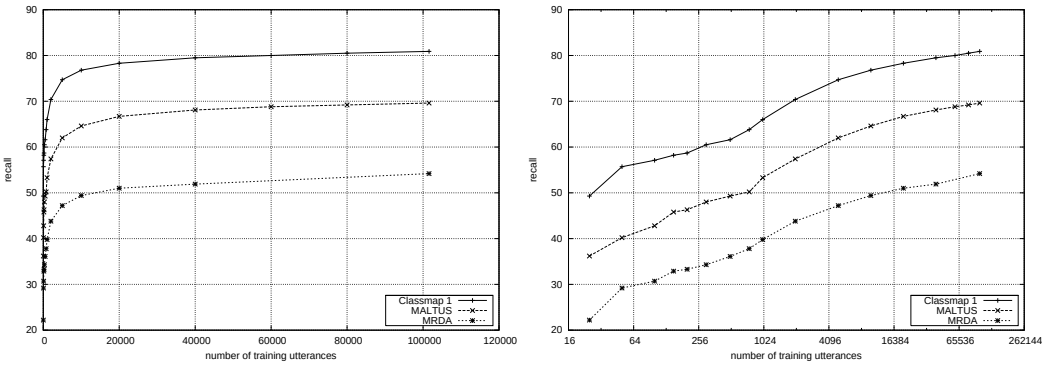

Figure 3: Recall (percent) for MRDA and MALTUS labels, and MRDA mapped with classmap 1, with different sizes of the training set. (linear and log scale, using 4-fold cross-validation, 2-fold for MRDA with 101584 training utterances)

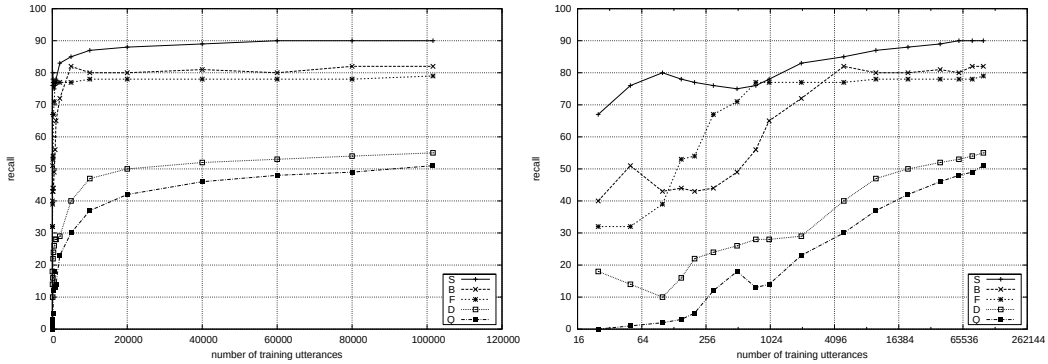

Figure 4: Recall (percent) for statements, questions, floors, backchannels and disruptions (classmap 1, linear and log scale, 4-fold cross-validation)

2.4 A New Metric for the Evaluation of Classification Results

Usually, classification tasks are evaluated using the precision and recall metrics:

$$Precision(l) := \frac{correct(l)}{tagged(l)}$$

$$Recall(l) := \frac{correct(l)}{occurs(l)}$$

where $occurs(l)$ is the number of times the label l occurs in the human annotation of the test corpus, $tagged(l)$ is the number of times it was assigned by the classifier, and $correct(l)$ is the number of times it was correctly assigned.

The recall values given in the experiments are the total recall over all labels:

$$Recall := \frac{\sum_l correct(l)}{\sum_l occurs(l)}$$

However, these are binary metrics which do not consider the case that the assigned label is incorrect, but very similar to the correct label. For instance, the label `s^rt` marks a statement with rising tone; we can hardly recognise this properly as we do not use phonological features. However, many such utterances will be tagged as `s` (statement). By defining a similarity metric between dialogue acts, we can include such cases in the evaluation of the classifier.

One way to define such a similarity metric is to order the labels in a hierarchy according to the sets of tags which make up the labels. For MRDA labels, this means we have several hierarchies with a general tag at the top (see fig. 1). Using such hierarchies, we can check if the "true" label and the classifier output have a least upper bound (lub). If there is one, there is at least some relationship between the labels. As we found in our experiments, in most cases where the lub exists, the classifier output is underspecific, i.e., some special tags are missing. Using this concept, we define a distance metric between two labels DA^T (a *true* label) and DA^C (a *classified* label):

$$\textsc{Scorre}(x,y) := \begin{cases} 1 - \frac{\delta^T + \delta^C}{2 \times depth} & \text{if } DA^{lub} \text{ exists} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$minPath(x,y) := \text{shortest path between x and y} \quad (2)$$

$$\delta^C := |minPath(DA^C, DA^{lub})| \quad (3)$$

$$\delta^T := |minPath(DA^T, DA^{lub})| \quad (4)$$

For our experiments with MRDA and MALTUS labels, we set $depth$ to 5 (with the current ordering of the labels in the ICSI corpus as shown in figure 1, the maximum distance between a *lub* and a label is 5); thus the denominator is 10, and a SCORRE of 0.9 means that the shortest path between two labels in the hierarchy has length 1.

For a test of a classifier with n utterances, true labels DA_i^T and classified labels DA_i^C, we define

$$\text{SCORRACY} = \frac{\sum_{i=1}^{n} \text{SCORRE}(DA_i^T, DA_i^C)}{n}$$

We motivate SCORRE by its similarity to *fScore* between two multi-dimensional labels (see also [8]). Considering labels as sets of tags (e.g. sˆrt as $\{s, rt\}$) allows us to define precision and recall for a true label DA^T and a classified label DA^C by using their intersection. Let

$$DA^I := DA^T \cap DA^C \quad (5)$$
$$\delta^C := |DA^C| - |DA^I| \quad (6)$$
$$\delta^T := |DA^T| - |DA^I| \quad (7)$$

For the normal labels in fig. 1, DA^I is equivalent to DA^{lub}, and the set-differences δ^T and δ^C are equivalent to the distances defined in (3) and (4). Now we can define *precision*, *recall* and *fScore* for a pair of labels DA^T and DA^C:

$$\begin{aligned} \textit{precision} &:= \frac{|DA^I|}{|DA^C|} = 1 - \frac{\delta^C}{|DA^C|} \\ \textit{recall} &:= \frac{|DA^I|}{|DA^T|} = 1 - \frac{\delta^T}{|DA^T|} \\ \textit{fScore} &:= \frac{2 * \textit{precision} * \textit{recall}}{\textit{precision} + \textit{recall}} \\ &= 1 - \frac{\delta^T + \delta^C}{|DA^T| + |DA^C|} \end{aligned}$$

Note the denominators: the distances are normalised to the sizes of the true and the classified labels. Conversely, SCORRE simply normalises to a constant chosen to ensure that it always yields a value between 1 and 0. Consequently, *precision*, *recall* and *fScore* determine which fraction of the output of a classifier is correct, while SCORRE and SCORRACY tell us how much it deviates from the ground truth.

In the following example, testing a classifier on 14512 utterances has resulted in 7823 correct and 4038 approximately correct classifications:

	utterances		∑Scorre	avg.
correct	7823	53.9%	7823	100%
approx.correct	4038	27.8%	3542.3	88%
all	14512	100%	11365.3	70%

Since each correct classification contributes 1 to the total SCORRE, and incorrect classifications do not contribute at all, the 4038 approximately correct classifications contribute 3542.3, or 88% on average, i. e., the average distance to the correct label in these cases is 1.

It is clear that this metric is highly dependent on the hierarchy of labels. Measuring the difference between labels by the length of the minimal path between them implies that we consider the edges in the hierarchy as representing equal differences between the content of labels. Without this assumption, one might introduce weights for the edges and define δ^C and δ^T as the sum of the weights on the cheapest path.

2.5 Classification Experiments

In this section, we report some classification experiments with the complex MRDA/MALTUS labels (that is, without regard to the internal structure of the labels), using an off-the-shelf maximum entropy classifier package for Java.[6]

A maxent model is trained from a set of examples, which consist of the features of an input utterance and its DA label (the class of the input). The resulting model maps $(feature, label)$ pairs to weights indicating how strongly the presence of $feature$ predicts $label$.

We used the following features:

- word features: the words occurring in the utterance, the initial and final words, and the initial words of the following utterance
- word bigrams: the bigrams occurring in the utterance, and the utterance-initial/final bigrams
- the length of the utterance
- temporal relation features indicating whether there is a pause, no pause or an overlap between the current utterance and the preceding/following one
- features indicating whether the current utterance is the beginning, or ending, or in the middle of a speaker turn
- the DA label of the preceding utterance

Note that some of these features are forward-looking. We would not want to use such features in a dialogue system which is required to react to a user's input; in a meeting-processing application, however, we can expect to be able to use at least the immediate context of an utterance. Note that we did not use any phonological features. Features, like stemming and part-of-speech information would be desirable.

We ran a series of classification experiments using the original MRDA labels, mapping the MRDA labels to MALTUS labels, and finally mapping the MRDA labels to the five categories "statement", "question", "backchannel", "floor management" and "disruptions" (the "classmap 1").

With MRDA and MALTUS labels, we find that only the most frequent labels occur frequently enough to be recognised reliably, or to have a significant influence on testing results.

Out of the 1256 MRDA labels, there are only 80 which occur more than 100 times. However, these 80 labels make up 111496 of all 118694 utterances (94%). There are 265 which occur 10 times or more. This means that about 80% of the labels occur only one to nine times; these labels are almost never correctly recognised. Table 3 shows results of one classification experiment: by simply using the labels as-is, we get approximately 51% correct classifications, and another 29% approximate classifications.

With MALTUS labels, we have significantly less labels (81), and their distribution over the corpus is less uneven: there are 23 labels which occur more than 100 times, and 42 which occur more than 10 times. When we train a classifier for these labels, we see that mostly those which occur more than 100 times are reliably recognised. Table 3 shows the results using the same training/testing set, but with the labels mapped to MALTUS labels. We can see that more utterances are correctly classified (67.1%) than with MRDA labels, and the sum of correct and approximately correct classifications is higher as well. (83.2%).

[4] reports a similar classification experiment without disruption marks and with a slightly different version of the MALTUS tag set and different features, achieving 73.2% accuracy.

The maximum generalisation of the tagset which can still be considered useful is to map all labels to one out of five classes: statements, questions, backchannels, floor management and disruptions. (Actually, there is a sixth class, "X" for non-speech noises. However, it is very rare.) We tried two variants of such a mapping:

[6] The Maximum Entropy Classifier by the Stanford NLP Department, available from `nlp.stanford.edu/downloads/classifier.shtml`

event type	MRDA	MALTUS
correct	51.0%	67.1%
overspecific	3.6%	2.7%
underspecific	19.2%	11.2%
neighbour	5.9%	2.1%
approx.correct	28.8%	16.1%
total	79.8%	83.2%

Table 3: Classification results using 20000 utterances as training material and 14512 for testing, 4-fold cross-validation

- One variant (the "classmap 1") comes with the documentation to the ICSI meeting corpus: this mapping prefers disruptions in some cases - for instance, a disrupted statement is mapped to D, not S. In this case, we only get a recall of 78.7%. A similar result—77.9%—was reported in [4].
- By mapping each label to one of the five classes according to its general tag, we have more instances of statements. The most frequent class which is recognised very well, with a recall of 91%. This leads to an increase of the total recall to 83.8%.
- For a four-way classification experiment—discriminating utterances between statements, questions, backchannels and floor management, and ignoring disruptions—[4] reports 84.9% correct classifications.

An algorithm for the Reduction of the Tagset

The uneven distribution of class frequencies has some disadvantages when we choose to model monolithic labels. The size of the model, and the time required to train it, are rather large, although most of the classes are almost never recognised. Therefore, we used the following approach to reduce the set of classes.

We define the entropy of a set of DA labels and an annotated corpus as

$$\begin{aligned} H &:= -\sum_{l \in labels} p(l) log_2 p(l) \\ p(l) &:= \frac{number\ of\ occurrences\ of\ l}{corpus\ size} \end{aligned}$$

and for a mother-daughter pair of DAs (m,d), the loss in entropy when d is mapped to m:

$$\begin{aligned} \Delta H(m,d) &:= p(m) log_2 p(m) + p(d) log_2 p(d) \\ &\quad -(p(m)+p(d)) log_2 (p(m)+p(d)) \end{aligned}$$

Then we find the pair (m,d) in the current set which minimises ΔH, and map all occurrences of d to m. This step is repeated until the set is reduced to a given size.

This method differs from simply choosing the n most frequent classes in that it considers collapses the selected pair (m,d) to m, no matter which one has the higher frequency (for instance, the label `qy^rt` occurs 1022 times, `qy` only 368 times). Also, the limitation to mother-daughter pairs means that the labels at the top of a hierarchy (e.g. `qy`) are never removed.

The most frequent classification error is that an instance of a more specific label (e.g., `s^bk`) is assigned a less specific label (`s`), which is counted as an approximately correct classification. When this pair is collapsed to the less specific one, the same classification would be considered correct. This is what happens when we go from MRDA to MALTUS labels, or even to the 5-way-mapping: we can see a shift from approximately correct to correct classifications, while the sum remains the same or improves slightly (in the range between 80% and 85%).

#das	correct	approx	total	SCORRACY
16	81.5%	0.0%	81.5%	82%
20	73.4%	8.2%	81.4%	81%
25	63.5%	17.7%	81.2%	79%
50	53.4%	27.1%	80.5%	77%
60	52.3%	28.0%	80.3%	77%
70	51.8%	28.4%	80.2%	77%
80	51.6%	28.6%	80.2%	77%
90	51.4%	28.7%	80.1%	76%
100	51.4%	28.8%	80.2%	76%
150	51.3%	28.8%	80.1%	76%
200	51.1%	29.0%	80.1%	76%
300	51.0%	29.1%	80.1%	76%
400	51.0%	28.9%	79.9%	76%
500	51.0%	29.0%	80.0%	76%
750	51.0%	29.0%	80.0%	76%

Table 4: Results (4-fold cross-validation) when the set of MRDA labels is simplified using the entropy-based mapping.

#das	correct	approx	total	SCORRACY
10	71.5%	11.9%	83.4%	82%
20	67.2%	16.1%	83.3%	81%
30	67.1%	16.2%	83.3%	81%
40	67.1%	16.2%	83.3%	81%
50	67.1%	16.1%	83.2%	81%
60	67.1%	16.1%	83.2%	81%
70	67.1%	16.1%	83.2%	81%
81	67.1%	16.1%	83.2%	81%

Table 5: Results (4-fold cross-validation) after mapping MRDA labels to MALTUS labels, and then simplifying using the entropy method. 81 is the full set of labels.

When we use the entropy-based method to define mappings to smaller subsets of the MRDA or MALTUS labels, we observe a similar effect; it only becomes visible when we reduce the set of labels to a very small size (e.g. 25 MRDA or 10 MALTUS labels). We also observe a small improvement in the SCORRE metric. We ascribe this to the uneven distribution of the the labels over the corpus. Therefore, this way of shrinking the set of labels does not seem very useful in improving the classification accuracy; however, it significantly reduces the time needed to train a classifier, and the space occupied by the model.

2.6 Discussion and Outlook

We have discussed the task of dialogue act classification for a multidimensional tag set. In particular, we have focussed on the MRDA tag set and the ICSI meeting corpus. We introduced a novel forgiving evaluation metric which utilises a hierarchical view of the tag set. The intuition behind SCORRE is that not hitting the correct tag can be viewed as more or less wrong. We thus depart from the monolithic view of classification results which has been used up until now, e.g., [96, 12].

We also presented a method to gradually reduce the tag set. We showed that, for our classifier, the overall

recognition rate does not change much unless the initial set of labels is reduced drastically, to 50 for the MRDA set, or 10 for MALTUS).

Future work includes the following topics:

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Sum
1	qr	1	.	.	.	.	.	.	.	.	6	.	1	.	.	.	.	1	.	.	.	9
2	**sˆaa**	.	**338**	.	.	**24**	.	.	4	40	62	.	.	.	.	.	.	12	**494**	.	.	974
3	qo	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	1
4	%	.	2	.	.	2	.	.	11	3	53	.	.	30	1	2	.	3	3	2	.	112
5	**sˆbk**	.	**89**	.	.	**412**	.	.	1	36	42	.	.	.	1	.	.	15	**287**	.	.	883
6	qh	1	.	.	.	.	4	.	.	.	26	.	5	.	.	.	.	5	.	9	.	50
7	x	.	.	.	.	.	.	.	.	.	7	.	2	1	.	.	.	.	2	.	.	12
8	fh	.	7	.	.	3	.	.	659	41	40	.	11	31	3	23	2	1	57	.	.	878
9	fg	.	70	.	.	28	.	.	72	105	16	.	1	14	3	.	.	.	21	.	.	330
10	s	1	54	.	.	29	3	.	7	7	6148	.	104	12	.	4	1	37	37	9	57	6510
11	qoˆrt	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	1
12	s.%–	1	.	.	.	.	.	.	1	.	340	.	102	2	.	1	1	3	.	.	3	454
13	%-	.	.	.	.	1	.	.	26	6	109	.	12	140	.	8	2	7	2	3	.	316
14	h	.	.	.	.	.	.	.	18	10	1	.	.	.	19	.	.	.	2	.	.	50
15	%–	.	1	.	.	.	.	.	26	3	59	.	23	39	.	29	4	.	.	.	.	184
16	qrr	.	.	.	.	.	.	.	.	.	13	.	.	3	.	.	16	1	.	1	.	34
17	qy	2	5	.	.	4	.	.	.	1	245	.	10	1	.	.	1	245	47	1	2	564
18	**b**	.	**78**	.	.	**89**	.	.	2	.	26	.	.	.	.	.	.	8	**2189**	1	.	2393
19	qw	.	.	.	.	.	.	.	.	.	47	.	4	1	.	.	.	2	2	97	.	153
20	sˆdf	.	.	.	.	.	.	.	.	.	447	.	9	.	.	.	.	3	.	1	144	604
	Sums	6	644	.	.	592	7	.	827	252	7688	.	284	274	27	67	27	344	3143	124	206	
	x=y	1	338	.	.	412	4	.	659	105	6148	.	102	140	19	29	16	245	2189	97	144	
	x≠y	5	306	.	.	180	3	.	168	147	1540	.	182	134	8	38	11	99	954	27	62	

Table 6: A confusion table for 20 MRDA tags. The labels in the rows are the correct labels, those in the columns are the classifier outputs. E.g., line 2 column 18 (494) means that **sˆaa** was misclassified as **b** 494 times—more often than it was correctly recognised.

Examining confusion matrices

In our classification experiments based merely on transcriptions of the ICSI meetings, there are some dialogue acts that are often mixed up. In the confusion matrix (table 6), we have highlighted three such dialogue acts: `sˆaa` (statement and accept), `sˆbk` (statement and acknowledgement) and `b` (backchannel). These acts are among the most frequently confused ones, and have been shown before to be hard to distinguish, e.g., [96]. This is partly because they share much of their vocabulary (“u-huh”, “yeah”, “right”, “okay”, “absolutely”...). To a degree, they can be distinguished by their acoustic and temporal properties. For instance, accepts and acknowledgements usually occur after another speaker has completed a phrase or utterance, while backchannels can occur in the middle of a phrase of another speaker.

When we find such a pair or group of easily confused labels, we should, on the one hand, try to compare the definitions of these labels, or the tags in them, in order to find new features which we can extract from our training data and which help discriminating between the labels. On the other hand, collapsing these acts would possibly enhance the quality of the classification as well, whereas such a decision has to be taken according to the requirements from the consumers of the classification.

Classifying aspects separately

In the experiments reported, we train a single classifier for complex labels which are actually combinations of tags representing different aspects of an utterance. This way, most of the rare combinations are nearly impossible to recognise.

A different approach would be to use several separate classifiers, one for each aspect of an utterance. For MRDA labels, we might use one classifier to decide on the general class of an utterance (statement, question, etc.), additional classifiers for groups of tags (e.g., to determine the type of a question), and binary classifiers to check for the presence of independent properties (e.g. rising tone). Using separate classifiers for the different aspects, we might be able to recognise rare combinations of tags more reliably; in particular, it would enable us to recognise combinations which did not occur in the training material.

On the other hand, however, we would lose information about correlations between tags which is included "for free" in a single classifier for the complex labels. In [4], a single classifier for complex MALTUS labels (which reached an accuracy of 73.2%) was compared to a combination of classifiers, which reached only 70.5%.

Feature analysis

The results in [4] were obtained by using roughly the same kinds of features as in this article—words, bigrams and features indicating the previous dialogue act and temporal overlap between utterances. Especially for words and bigrams, further research is necessary, as their number is almost unlimited. It may prove worthwhile to further investigate to which degree different features add to the overall recognition result. Not only is the memory needed to store these features reduced, the same argument also applies to the time needed to train the classifier. One preliminary result is that ignoring words and bigrams with low frequencies (< 10) has almost no influence on the classification results.

Adding features

The features we use currently are those which are easy to obtain from the transcriptions available to us; however, they are suboptimal for recognising certain types of utterances. As fig. 4 shows, questions are the type with the worst recall, and we expect an improvement if phonological features were included. Also, we would like to include part-of-speech information.

Improving the modelling

Although our classifier evaluation takes similarities between labels into account, the maxent classifier package does not. The training procedure classifies the training data according to the current feature weights and adjusts the weights to minimise an error function. This function is based on the number of incorrect classifications and does not recognise partly correct ones. We are going to research whether the quality of the models can be improved by using an error function which is aware of similarities between labels.

References

[1] Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Michael Kipp, Stephan Koch, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. Dialogue Acts in VERBMOBIL-2 – Second Edition. Verbmobil-Report 226, DFKI Saarbrücken, Universität Stuttgart, Technische Universität Berlin, Universität des Saarlandes, 1998.

[2] James Allen and Marc Core. Draft of DAMSL: Dialog Act Markup in Several Layers. `http://www.cs.rochester.edu/research/ cisd/ resources/ damsl/ RevisbedManual/ RevisedManual.html`, 1997.

[3] J. L. Austin. *How to do Things with Words*. Oxford University Press, 1962.

[4] A. Clark and A. Popescu-Belis. Multi-level dialogue act tags. In *Proceedings of SIGDIAL '04 (5 th SIGDIAL Workshop on Discourse and Dialog)*, Cambridge, MA., 2004.

[5] Debra Biasca Daniel Jurafsky, Elizabeth Shriberg. Switchboard swbd-damsl shallow-discourse-function annotation (coders manual, draft 13). Technical report, University of Colorado, Institute of Cognitive Science, feb 1997. http:// www.colorado.edu/ linguistics/ faculty/ jurafsky/ pubs.html#Tech.

[6] Rajdip Dhillon, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. Meeting recorder project: Dialog act labeling guide. Technical report, International Computer Science Insitute, February 2004. ICSI Technical Report TR-04-002.

[7] 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 ICASSP-2003, Hong Kong*, Hong Kong, 2003 2003. ICASSP.

[8] Stephan Lesch, Thomas Kleinbauer, and Jan Alexandersson. "a new metric for the evaluation of dialog act classification". In *"Proceedings of Dialor05, the Ninth Workshop On The Semantics And Pragmatics Of Dialogue (SEMDIAL)"*, 2005.

[9] Andrei Popescu-Belis. Dialogue act tagsets for meeting understanding: an abstraction based on the damsl, switchboard and icsi-mr tagsets. Technical report, ISSCO/TIM/ETI, University of Geneva, September 2003. Version 1.2 (December 2004).

[96] Norbert Reithinger and Martin Klesen. Dialogue Act Classification Using Language Models. In *Proceedings of the* 5^{rd} *European Conference on Speech Communication and Technology (EUROSPEECH-97)*, pages 2235–2238, Rhodes, 1997.

[11] John R. Searle. *Speech Acts*. University Press, Cambridge, GB, 1969.

[12] 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, 2000.

3 Addressing

In this section we present preliminary results on automatic predication of the addressee of dialogue acts in four participants face-to-face meetings using Bayesian Networks and Naive Bayes classifiers. For training and testing classifiers, we have developed a small multi-modal corpus of hand-annotated meeting dialogues. The corpus contains several meetings from the M4 and the AMI pilot data collections. Due to the limitation of the available amount of training and testing data, our focus in the experiments presented in this section is not to build a high performance classifier. The focus is to find appropriate models for addressee classification that can be applied on the large AMI data set. Our goals are (1) to find relevant features for addressee classification using information obtained from multi-modal resources, (2) to explore to what extent the performances of addressee classifiers can be improved by combining different types of features and (3) to compare the performances of the Bayesian Network classifier and the Naive Bayes classifier for the task of addressee prediction.

3.1 Addressees and addressing behavior

When speakers design their utterances they assign different hearers to different roles. Goffman [45] distinguished three basic kinds of hearers to talk: those who overhear, whether or not their unratified participation is unintentional and whether or not it has been encouraged; those who are ratified but are not *specifically* addressed by the speaker (also called "unaddressed" recipients Goffman [47] or side-participants Clark and Carlson [31]); and those ratified participants who are addressed. Ratified participants are participants that are allowed to take part in conversation, that "have declared themselves open to one another for purposes of spoken communication and guarantee together to maintain a flow of words" [46].

Goffman [45] defined *addressees* as those "ratified participants () oriented to by the speaker in a manner to suggest that his words are particularly for them, and that some answer is therefore anticipated from them, more so than from the other ratified participants". According to this, it is the speaker who selects his addressee; the addressee is the one who is expected by the speaker to respond, who is invited by the speaker to take the floor. Addressing is a form of orienting or directing performed by a speaker.

In meeting conversations, a speaker may address his utterance to the whole group of participants present in the meeting, or to a particular subgroup of them, or just to one single participant in particular. Sometimes the speaker just thinks aloud or mumbles to himself without really addressing anybody. Examples of self addressed speech include utterances such as "Oops!" (after spilling water on the table) or "What else do I want to say?" (while trying to evoke more details about the issue that he is presenting). We excluded self-addressed speech from our study. In a group discussion, many of the dialogue acts are simply addressed to the group as a whole. However, when a speaker shows by verbal or non-verbal behavior that he intends to affect one selected participant or a subgroup of participants in particular, and to whom therefore he is giving primary attention in the present act then we see that participant or that group of participants as the addressee of the dialogue act that the speaker performs. The Goffman's definition cited above fits the initiatives like questioning or suggesting rather than the responses like answering or accepting. In our definition of addressing, a *direct* response to a request of a previous speaker who requested some information or opinion to be provided primarily to him is addressed to that speaker. Sometimes the questioner may request from the current speaker that he provides a response to the group (e.g. *What do you think about current proposal?*). Responses to these types of requests are mostly addressed to the group. In some cases, the speaker can understand a previous request as a stimulus to say something more i.e. to clarify, elaborate or explain the raised issue, addressing the whole group.

In conversations involving more than two people, most utterances are intended to be understood not only by the people being addressed but also by others. By saying "What do you think of this John?" the speaker not only addresses his question to John, he also *informs* all hearers about the act that he is simultaneously performing towards John. According to Clark and Carlson [31], the speaker performs two acts with each utterance in conversation involving more than two people. One is the traditional kind directed to the addressee (addressee-directed illocutionary act), and the other, called *informative* is directed to all ratified participants in the conversation (participant-directed illocutionary act). All addressee-directed acts are performed by means of informatives [31]. Consider the

following example:

A to B in front of C Did you like the book?

B to A in front of C yes, very much

A to C in front of B *and you?*

C to A in front of B I liked it too

When A asks B his question, A must also be informing C what he is asking B, otherwise A cannot be certain that C will understand the question *and you?*.

When a speaker A addresses B saying "I think John can tell you this" while it is known by the speaker that John is present and listening, we say that John is *indirectly addressed* by speaker A. The deictically used second person pronoun 'you' refers to B, the addressee of A's utterance.

Addressing behavior is behavior that speakers show to express to whom they are addressing their speech. It depends on the course of the conversation, the status of attention of participants, their current involvement in the discussion as well as on what the participants know about each others' roles and knowledge, whether explicit addressing behavior is asked for. If the speaker knows that his addressee is already attentive to what he says he does not have to call his attention. Using a vocative is the explicit verbal way to address someone. Addressees can be designated in partially by gestures. In some cases the speaker identifies the addressee of his speech by looking at the addressee, sometimes accompanying this by deictic hand gestures. A speaker can also exclude certain people as addresses by turning his back to them. Addresses can also be designated by the manner of speaking. For example, by whispering, a speaker can select a single individual or a group of people as addresses, letting everyone else know that they are not addressed. Addresses are often designated by the content of what is being said. For example, *"We have to decide together about the conceptual design"* is a suggestion addressed to the whole group.

3.2 Gaze behavior and addressing

Most studies into the function of gaze behavior in conversational interaction were based on dyadic conversations. Analyzing dyadic conversations, researchers in conversational analysis observed that gaze in social interaction is used for several purposes: to control communication (e.g. turn-taking), to provide feedback on the reaction of others, to communicate the nature of relationships (e.g. dominant relationship or dependent relationship), to communicate emotions and to avoid distraction through avoiding excess input of information [64, 8].

Recent studies into multi-party interaction, addressed the question about functions of gaze in addressing behavior. [120] investigated to what extent the focus of visual attention might function as an indicator for the focus of "dialogic attention" in four-participants face-to-face conversations. "Dialogic attention " includes attention while listening to a person as well as attention while talking to one or more persons. The empirical findings show that when a speaker is addressing an individual, there is 77% chance that the gazed person is the addressed individual. When addressing a triad, speaker gaze seems to be evenly distributed over listeners in the situation where conversational participants are seated around the table. It is also shown that on average a speaker spends significantly more time gazing at an individual when addressing the whole group, than at others when addressing a single individual. When addressing an individual, people gaze 1.6 time more while listening (62%) than while speaking (40%). In the situation when a triad is addressed, the amount of speaker gaze increases significantly to 59%. According to all these estimates, we can expect that gaze directional cues are good indicators for addressee prediction.

However, these findings cannot be generalized in the situations when some objects of interests are present in the conversational environment, since it is expected that the amount of time spent looking at the persons will decrease significantly. As shown in [11], in a situation when a user interacts with a multimodal information system and in the meantime talks to another person, the user looks most of the time at the screen, both when talking to the system (94%) and when talking to the user (57%). Also, another person looks at the system in 60% of cases when talking to the user. This indicates, that gaze is a less powerful cue for addressee predication in the situation when objects

of interests are present in the environment. [11] also showed that some improvement in addressee detection can be achieved by combining utterance duration with facial orientation.

In meeting conversations, the contribution of the gaze direction for addressee prediction is also dependent on the current meeting activity. For example, when giving a presentation, a speaker most probably addresses his speech to the whole audience, although he may only look at a single individual in the audience.

Since it is very difficult to record eye gazing of meeting participants, the information about visual focus of attention can be automatically induced from head orientation [106, 63].

We explored not only the effectiveness of the speaker's gaze direction, but also the effectiveness of the listeners' gaze directions as cues for addressee prediction in two situations: (1) when using only gaze to identify who is addressed and (2) when combining gaze information with other sources of information.

3.3 Data collection

To train and test addressee classifiers, we developed a small corpus of hand-annotated meeting dialogues. The meetings were recorded in the IDIAP meeting room in the research program of the M4 and AMI projects (AMI pilot meetings). The corpus contains hand-annotated dialogue acts, adjacency pairs, addressees and gaze directions of meeting participants. Each type of annotation is described in detail in [59].

Our dialogue act tag set is based on the MRDA (Meeting Recorder Dialogue Act) set [36]. It is a MRDA "classmap", made by grouping the MRDA tags into 17 categories. In contrast to MRDA, where each functional utterance is marked with a label compound of one or more tags from the set, each functional utterance in our DA schema is marked as Unlabeled or labeled with exactly one tag from the set that is presented in Table 7.

DA tag set	**MRDA**
Statements	
s Statement	s Statement
Questions	
q Information-Request	Wh-question, Y/N question, OR-question, Or Clause After Y/N question
qo Open-ended Question	Open-ended questions
qh Rhetorical Question	Rhetorical Questions
Backchannels and Ack.	
bk Acknowledgement	Acknowledgment,Backchannel
ba Assessment/Appreciation	Assessment/Appreciation
Responses	
rp Positive response	(Partial)Accept, Affirmative Answer
rn Negative response	(Partial)Reject, Dispreferred and Negative Answer
ru Uncertain response	Maybe , No Knowledge
Action Motivators	
al Influencing-listeners-action	Command, Suggestion
as Committing-speaker-action	Commitment, *Suggestion*
Checks	
f "Follow Me"	"Follow Me"
br Repetition Request	Repetition Request
bu Understanding Check	Understanding Check
Politeness Mechanisms	
fa Apology	Apology
ft Thanks	Thanks
fo Other polite	Downplayer,Sympathy, Welcome

Table 7: Dialogue act tag set

Labelling of adjacency pairs consists of marking dialogue acts that occur as their a-part and b-part. If a dialogue act is an a-part with several b-parts, for each of these b-parts, a new adjacency pair is created.

Since all meetings in the corpus consist of four participants, addressee of a dialogue act is labeled as *Unknown* or with one of the following addressee tags: individual P_x, a subgroup of participants P_x, P_y or the whole audience P_x, P_y, P_x.

Labeling gaze direction denotes labeling gazed targets for each meeting participants. For addressee identification, the only targets of interests are meeting participants. Therefore, the tag set contains tags that are linked to each participant (P_x) and the *NoTarget* tag that is used when the speaker does not look at any of the participants.

Annotators involved in the corpus design were able to reproduce the gaze annotation reliably (segmentation 80.40% (N=939); classification $\kappa = 0.95$). Annotators involved in dialogue act, adjacency pairs and addressee annotations were divided into two groups; each group annotated different sets of meeting data. Table 8 shows reliability of dialogue act segmentation as well as Kappa values for dialogue act classification and addressee annotation for each annotation group.

Group	Segmentation (%)	N	DA(κ)	ADD(κ)
B&E	91.73	377	0.77	0.81
M&R	86.14	367	0.70	0.70

Table 8: Inter-annotator agreement on DA and addressee annotation: N - the number of agreed segments

3.4 Addressee classification

This section presents preliminary results on addressee classification in four-persons face-to-face conversations using Bayesian Network and Naive Bayes classifiers.

In a dialogue situation, which is an event that lasts as long as the dialogue act performed by the speaker in that situation, the class variable is the addressee of the dialogue act performed by the current speaker (ADD). Since there are only few instances of subgroup addressing present in the data, we removed them from the data set and excluded all possible subgroups of meeting participants from the set of class values . Therefore, we define addressee classifiers to identify one of the following class values: individual P_x where $x \in \{0,1,2,3\}$ and *ALLP* which denotes the whole audience.

3.4.1 Features

To identify the addressee of a dialogue act, we used three sorts of features: contextual features, utterance features and gaze features.

Contextual features provide the information about the preceding utterances. We experimented with using the information about speaker, addressee and dialogue act of the immediately preceding utterance on the same or a different channel (SP-1, ADD-1, DA-1) as well as the information about the related utterance (SP-R, ADD-R, and DA-R). A related utterance is the utterance that is the A part of an adjacency pair with the current utterance as the B part. The information about the speaker of the current utterance (SP) has also been included in the contextual feature set.

As utterance features, we used a set of lexical features that are based on our intuition on which words are the most informative for indicating whether the utterance is single or group addressed. The set includes the following features:

- does the utterance contain personal pronouns we or you, both of them, or neither of them?
- does the utterance contain possessive pronouns or possessive adjectives (your/yours or our/ours), their combination or neither of them?
- does the utterance contain indefinite pronouns such as somebody, someone, anybody, anyone, everybody and everyone?
- does the utterance contain the name of speaker P_x?

Utterance features include also the information about the utterance's conversational function (DA tag) and the information about utterance duration i.e. whether the utterance is short or long. In our experiments, an utterance is considered as a short utterance, if its duration is less or equal to 1 sec.

We experimented with a variety of gaze features. In the first experiment, for each participant (P_x) we defined a set of features in the form $P_x - looks - P_y$ and $P_x - looks - NT$ where $x, y \in \{0,1,2,3\}$ and $x \neq y$; $P_x - looks - NT$ represents that participant P_x does not look at any of the participants. The value set represents the number of times that speaker P_x looks at P_y or looks away during the time span of the utterance: "zero" for 0, "one" for 1, "two" for 2 and "more" for 3 or more times. In the second experiment, we defined a feature set that incorporates only the information about gaze direction of the current speaker ($SP - looks - P_x$ and $SP - looks - NT$) with the same value set as in the first experiment.

3.4.2 Results

To train and test the Bayesian Network and Naive Bayes classifiers, we used the hand-annotated M4 data. After we had discarded the instances labeled with *Unknown* or subgroup addressee tags, 781 instances left available for the experiments.

For learning the Bayesian network structure, we applied the K2 algorithm [33]. The algorithm requires an ordering on the observable features; different ordering leads to different networks structures. We conducted experimenters with several orderings. The obtained classification results for different orderings were nearly identical. For learning conditional probability distributions, we used the algorithm implemented in the WEKA toolbox[7] that produces direct estimates of the conditional probabilities.

The performances of the classifiers are measured using different features sets. First, we measured the performances of classifiers using utterance features, gaze features and contextual features separately. Then, we conducted experiments with all possible combinations of different types of features. For each classifier, we performed 10-fold cross-validation. Table 9 summarizes the accuracies of the classifiers for different feature sets (1) using the gaze information of all meeting participants and (2) using only the information about speaker gaze direction (with 95% confidence interval).

	BN		NB	
Feature sets	Gaze All	Gaze SP	Gaze All	Gaze SP
All Features	81.05%(±2.75)	82.59%(±2.66)	78.10%(±2.90)	78.49%(±2.88)
Context	73.11%(±3.11)		68.12%(±3.27)	
Utterance+SP	52.62%(±3.50)		52.50%(±3.50)	
Gaze+SP	66.45%(±3.31)	62.36%(±3.40)	64.53%(±3.36)	59.02%(±3.45)
Gaze+SP+Short	67.73%(±3.28)	66.45%(±3.31)	65.94%(±3.32)	61.46%(±3.41)
Context+Utterance	76.82%(±2.96)		72.21%(±3.14)	
Context+Gaze	79.00%(±2.86)	80.03%(±2.80)	74.90%(±3.04)	77.59%(±2.92)
Utterance+Gaze+SP	70.68%(±3.19)	70.04%(±3.21)	69.78%(±3.22)	68.63%(±3.25)

Table 9: Classification results for Bayesian Network and Naive Bayes classifiers using gaze information of all meeting participants(Gaze All) and using speaker gaze information (Gaze SP)

The results show that the Bayesian Network classifier outperforms the Naive Bayes classifier for all feature sets, although the difference is significant only for the feature sets that include contextual features.

For the feature set that contains only the information about gaze behavior combined with the information about the speaker (Gaze+SP), both classifiers perform significantly better when exploiting the gaze information of all meeting participants. In other words, when using solely the focus of visual attention to identify the addressee of a dialogue act, the focus of attention of non-speaking participants provides valuable information for addressee prediction. The same conclusion can be drawn when adding the information about utterance duration to the gaze feature set (Gaze+SP+Short), although for the Bayesian Network classifier the difference is not significant. For all other feature sets, the classifiers do not perform significantly different when including or excluding the listeners

[7] http://www.cs.waikato.ac.nz/ ml/weka/

gaze information. Even more, both classifiers perform better using only the speaker gaze information in all cases except when combined utterance features and gaze features are exploited (Utterance+Gaze+SP).

The Bayesian network and Naive Bayes classifiers show the same changes in the performances over different feature sets. The results indicate that the selected utterance features are less informative for addressee prediction (52.50%) compared to contextual features (BN:73.11%; NB:68.12%) or features of gaze behavior (BN:66.45%, NB:64.53%). The results also show that adding the information about the utterance duration to the gaze features, slightly increases the accuracies of the classifiers (BN:67.73%, NB:65.94%), which confirms findings presented in [11]. Combining the information from the gaze and speech communication channels improves significantly the performances of the classifiers (BN:70.68%; NB:69.78%) in comparison to performances obtained from each channel separately. Furthermore, higher accuracies are gained when adding contextual features to the utterance features (BN:76.82%; NB:72.21%) and to the features of gaze behavior (BN:80.03%, NB:77.59%). As it is expected, the best performances are achieved by combining all three types of features (BN:82.59%, NB:78.49%), although not significantly better compared to combined contextual and gaze features.

We also explored how well the addressee can be predicted excluding information about the related utterance (i.e. AP information). The best performances are achieved using speaker gaze information in combination with contextual and utterance features (BN: 79.39%; NB: 76.06%). A small decrease in the classification accuracies when excluding AP information (about 3%) indicates that remaining contextual features, utterance features and gaze features capture most of the useful information provided by AP.

3.4.3 Evaluation of the addressee classifiers on the AMI pilot data

We investigated how well the classifiers trained on the M4 data perform on the AMI pilot data. Two AMI pilot meetings were used for the evaluation, although only one of them has been annotated with visual focus of attention. After discarding utterances labeled with *Unknown* and subgroup addressee tags from the data set, we had 291 instances available for testing the performances of the classifiers using the complete feature set and 673 instances for testing the performances of classifiers using combined contextual and utterance features.

The results presented in Table 10 show a significant decrease in the performances of classifiers for both feature sets in comparison to the performances on the M4 data. The accuracies decrease about 10% in all cases, except for the Naive Bayes classifier when visual information is used (more than 13%).

	BN		NB	
Feature sets	Gaze All	Gaze SP	Gaze All	Gaze SP
All Features	70.65%	72.35%	63.14%	65.87%
Context+Utterance	66.41%		63.45%	

Table 10: Classification results for the M4 classifiers on the AMI pilot data

This decrease in the performances can be due to several reasons. First, there are more single-addressed utterances in the AMI meetings than in the M4 meetings. Second, single-addressed utterances in the M4 meetings are almost equally distributed over all participants, whereas in the AMI meetings the distribution is dependent on the roles participants play in a meeting: a participant with the dominate role (i.e. project manager) has been addressed more than the others (40.19%). Third, participants in the AMI meetings show different gaze behavior, since their attention is focused part of the time at the task object i.e. the remote TV control that is present in the meeting room, especially when the remote control is relevant for the topic of conversation. As discussed in Section 3.2, the presence of the object of interest decreases the effectiveness of the gaze as an indicator of who is being addressed.

From this we can conclude, that including the background knowledge about participants' roles in a meeting as well as the information about the topic of conversation may improve addressee prediction on the AMI data.

4 Dominance detection

In many cases it is beneficial for the effectiveness of a meeting if people assume a cooperative stance. Grice [50] formulated four maxims that hold for cooperative conversations. The maxims of quantity, quality, relevance and manner state that one should say nothing more or less than is required, speak the truth or say only things for which one has enough evidence, only say things that are relevant for the discussion at hand and formulate the contribution such that it can be easily heard and understood by the interlocutors. These maxims are all formulated from the perspective of producing utterances in a conversation. One could define similar maxims for cooperative behavior, more generally. One can also think of several tasks of chairpersons in meetings as being guided by such maxims. The chair should facilitate the participants to have their say, to cut off people who make their contribution too long or to intervene when contributions are not relevant to the discussion at hand. Discussions should be properly organized to have arguments develop, so that all positions are put to the fore, and all relevant pros and cons are raised. People that are too dominant in meetings may violate one or more of the cooperative maxims and are thereby frustrate the process of collective decision making for which many meetings are intended. The chair of the meeting should avoid this from happening or intervene when it does.

Nowadays, in order to maximize the efficiency, meetings can be assisted with a variety of tools and supporting technologies [97]. These tools can be passive objects such as microphones facilitating better understanding or semi-intelligent software systems that automatically adjust the lighting conditions. In the near future, meetings will be assisted with various similar sorts of active, and perhaps even autonomous, software agents that can make sense of what is happening in the meeting and make certain interventions [41]. An example of such meeting assisting agents could be an agent that signals possible violations of cooperative maxims in the decision making process to the chairperson. One of the major issues to be addressed in this case is how the agent can detect that there is such a disturbance.

4.1 Dominance

According to Hoffmann [56], there are three types of behavioral roles that can be identified in groups or teams. These roles can be classified as task-oriented, relation-oriented and self-oriented. Each group member has the potential of performing all of these roles over time. *Initiators, Coordinators* and *Information Givers* are task-oriented roles that facilitate and coordinate the decision making tasks. The Relations-Oriented role of members deals with team-centered tasks, sentiments and viewpoints. Typical examples are : *Harmonizers, Gatekeepers* and *Followers*. The Self-Oriented role of members focusses on the members' individual needs, possibly at expense of the team or group. Examples here are *Blockers, Recognition Seekers* and *Dominators*. The Dominator is a group member trying to assert authority by manipulating the group or certain individuals in the group. Dominators may use flattery or proclaim their superior status to gain attention and interrupt contributions of others. According to Hellriegel et al. [55], a group dominated by individuals who are performing self-oriented sub-roles is likely to be ineffective.

In psychology, dominance refers to a social control aspect of interaction. It involves the ability to influence others. One can refer to it as a personality characteristic - the predisposition to attempt to influence others - or one can use the term to describe relationships within a group. Dominance is a hypothetical construct that is not directly observable. However, there appear to be certain behavioral features displayed by people that behave dominantly that make it possible for observers of these behaviors to agree on judgments of dominance. In Ellyson and Dovidio [42] the nonverbal behaviors that are typically associated with dominance and power are investigated. In several of the papers in that volume, human perceptions of dominance are discussed as well.

In "A System for Multiple Level Observation of Groups" (SYMLOG), [13], Bales distinguishes three structural dimensions in group interactions: status, attraction and goal orientation. Goal orientation refers to the way people are involved with the task or rather with socio-emotional behaviours. This dimension was already present in Bales' earlier work on Interaction Process Analysis [12]. The attraction dimension concerns friendly versus unfriendly behaviours. The status dimension has to do with dominant versus submissive behaviours. Bales developed a checklist that observers can use to structure their observations of groups. He has also developed a number of self-

report scales that group members can use to rate themselves (and other group members) on these three dimensions. SYMLOG presents a questionnaire containing 26 questions from which 18 relate to the concept of dominance. The factors involved in these questions provide a frame for the meaning of the concept. An overview of these factors in their most general form are shown in Table 11.

Positive contributions	Negative contributions
active, dominant, talks a lot	passive, introverted, said little
extraverted, outgoing, positive	gentle, willing to accept responsibility
purposeful, democratic task-leader	obedient, worked submissively
assertive, business-like, manager	self-punishing, worked too hard
authority, controlling, critical	depressed, sad, resentful, rejecting
domineering, tough-minded, powerful	alienated, quit, withdrawn
provocative, egocentric, showed-off	afraid to try, doubts own ability
joked around, expressive, dramatic	quietly happy just to be in group
entertaining, sociable, smiled, warm	looked up to others, appreciative

Table 11: Aspects of dominance according to SYMLOG

When we look at this scale we see that it is very hard to operationalize many factors - such as 'purposeful' and 'alienated', for instance. They depend on human interpretative skills. What we need are automatically detectable features that can be quantified and transformed as a series of digits into our system.

To train a classifier that can determine who is the person that dominated a meeting, we need a corpus of meeting recordings with the relevant features that the classifier is using either extracted or annotated and also we need to know how the participants of the various meetings scored on the dimension of dominance. We will provide more details on the corpus and the features used by the classifier in Section 4.3. Now, we will first describe how we established the dominance ranking for the meetings we used.

4.2 Dominance judgements

We used a corpus of eight four-person meetings[8]. The meetings varied in length between 5 and 35 minutes. We collected 95 minutes in total. We used different kinds of meetings, including group discussions where statements had to be debated, discussions about the design of a remote control, book club meetings and PhD. evaluation sessions.

We asked ten people to rank the participants of the meetings. Each person ranked four, i.e. half of, the meetings. We thus had a total of five rankings for every meeting. We simply told people to rate the four people involved in the meeting on a dominance scale. We did not tell the judges anything more about what we meant by that term. The results are shown in Table 12. The first cell shows that in the first meeting (M1), judge A1 thought that the most dominant person was the one corresponding to the fourth position in this list, second was the first person in this list, third the second person in the list and least dominant was the third person in the list: 2,3,4,1. If one looks at the judgements by the other judges for this meeting (A2 to A5), by comparing the different columns for this first row, one can see that A3's judgments are identical to A1's. All but A4 agree that the fourth person on the list was most dominant. All but A5 agree that the third person was least dominant. All but A2 agree that the first person was the second dominant person. This seems to suggest that on the whole judgements were largely consistent across judges.

To establish the degree of agreement, we compared the variance of the judgements with the variance of random rankings. If the variance of the annotators is smaller than the variance of the random rankings, we have a strong indication that people agree on how to create a dominance ranking.

[8]Five of these were recorded for the M4 project (M4TRN1, M4TRN2, M4TRN6, M4TRN7 and M4TRN12) and three for the AMI project, two of them were pilot meetings (AMI-Pilot 2 and AMI-Pilot 4) and the third one was a meeting from the AMI spokes corpus (AMI-FOB 6).

	A1	A2	A3	A4	A5	'Average'	'Variance'
M1	2,3,4,1	3,2,4,1	2,3,4,1	2,1,4,3	2,4,3,1	2,3,4,1	8
M2	2,3,4,1	2,3,4,1	2,3,4,1	2,3,1,4	3,2,4,1	2,3,4,1	8
M3	2,1,3,4	3,1,2,4	2,1,4,3	3,1,2,4	1,2,3,4	2,1,3,4	8
M4	2,4,3,1	2,4,3,1	1,4,2,3	2,3,4,1	1,4,3,2	1,4,3,1	4
	A6	A7	A8	A9	A10	'Average'	'Variance'
M5	4,3,2,1	4,3,1,2	3,4,1,2	4,3,1,2	3,4,1,2	4,3,1,2	6
M6	1,3,2,4	1,4,3,2	3,1,4,2	3,1,4,2	1,3,4,2	1,3,4,2	12
M7	1,4,3,2	2,4,3,1	3,2,1,4	2,4,1,3	1,4,3,2	1,4,2,3	14
M8	1,2,4,3	1,4,2,3	2,1,3,4	2,1,3,4	1,2,4,3	1,2,3,4	12

Table 12: Rating of meeting participants for all the annotators per meeting.

If we add up the dominance scores for each person in the meeting, this results for the first meeting in scores 11, 13, 19 and 7, with results in an overall ranking of 2, 3, 4, 1. We call this the 'average' ranking. In case of similar scores, we scored them an equal rank, letting the other two ranks behind. For each of the judges we compare how they differ for each person from this average.

As a measure for the variance we calculated the sum of all the (absolute) differences of each of the annotators judgments (A^i) with their corresponding average. The difference with the average was calculated as the sum of the pairwise absolute differences for all the annotators values of the meeting participants A_p with their corresponding average value $Average_p$. See Table 12 for the results.

$$'Variance' = \sum_{i=1}^{5} \sum_{p=1}^{4} |A_p^i - Average_p|$$

In this case A1 and A3 judgments are identical to the average. A2 made different judgments for the first person (scoring him as 3 instead of 2) and the second person (scoring him as 2 instead of 3). So this results in a variance of 2 adding up the variance 4 and 2 of judges A4 and A5 respectively this ends up in an overall variance of 8 for judgements on the first meeting.

When comparing the variance of the judges with the variance resulting from randomly generated rankings, the distribution of the variance of the annotators (μ = 9, σ = 3.38, $n = 8$) lies far more left of the distribution coming from randomly generated rankings. (μ = 17.8, σ = 3.49, $n = 1.0 * 10^6$). The two distributions appeared to be statistically significant ($p < 0.001$) according to the 2-sided Kolmogorov Smirnov test. It thus appears that judges agree very well on dominance rankings. We may have to be conservative to generalize this though as we have only a small (n=8) amount of real samples.

These results support our initial thoughts, where we expected humans to agree (to a reasonable extent) on the ranking of meeting participants according to their conveyed dominance level.

4.3 Features used by the classifier

Dominance can be regarded as a higher level concept that can may be deduced automatically from a subset of lower level observations ([93]), similar to the assignment of the value for dominance by humans on the basis of the perception and interpretation of certain behaviours.

For our classifier we considered some common sense features that possibly could tell us something about the dominance of a person in relation to other persons in meetings. For each person in the meeting we calculated scores for the following features.

- The person's influence diffusion (IDM)
- The speaking time in seconds (STS)

- The number of turns in a meeting (NOT)
- The number of times addressed (NTA)
- The number of successful interruptions (NSI)
- The number of times the floor is grabbed by a participant (NOF)
- The number of questions asked (NQA)
- The number of times interrupted (NTI)
- The ratio of NSI/NTI (TIR)
- Normalised IDM by the amount of words spoken. (NIDF)
- The number of words spoken in the whole meeting (NOW)
- The number of times privately addressed (NPA)

The *Influence diffusion model* [81] generates a ranking of the participants by counting the number of terms, reused by the next speaker from the current speaker. The person who's terms are re-used the most is called the most influential.

Most of the features appear as simple metrics with variations that measure the amount to which someone is involved in the conversation and how others allow him/her to be involved. These are all measures that are easy to calculate given a corpus with appropriate transcriptions and annotations provided. Metrics used in the literature, as in SYMLOG, depend on the interpretation of an observer.

After the judges that rated our corpus had finished their ratings, we asked them to write down a list of at least five aspects which they thought they had based their rankings on.

> Dominant is the person: who speaks for the longest time, who speaks the most, who is addressed the most, who interrupts the others the most, who grabs the floor the most, who asks the most questions, who speaks the loudest, whose posture is dominant, who has the biggest impact on the discussion, who appears to be most certain of himself, who shows charisma, who seems most confident.

From the features identified by the annotators we can see that e.g. *charisma* and *confidence* are again typical examples of features that are very hard to measure and to operationalize. Most of this is again due to the fact that a proper scale does not exist, and as a result the valuation becomes too subjective and values from one annotator might not correlate with the values from another annotator. Several of these features are similar to the ones we are exploring for their predictive power in our classifier.

4.4 Acquiring and preprocessing the data

For each of the eight meetings ranked by our annotators, we calculated the values for the measures identified in the previous section. This was done on the basis of simple calculations on manual annotations and on the results of some scripts processing the transcriptions[9]. With respect to addressee annotation 25% of the data was not annotated due to the cost involved[10].

In order to make the values for the same feature comparable, we first made the feature values relative with respect to the meeting length. This was done in two steps. First the fraction, or share, of a feature value was calculated given all the values for that feature in a meeting.

[9] All transcriptions used were created using the official AMI and M4 transcription guidelines of those meetings [76, 39].
[10] Addressee information takes over 15 times real time to annotate [59].

$$\text{The share of a feature value } (F'_{Pn}) = \frac{F_{Pn}}{\sum F_{P1}..F_{P4}}$$

Then, according to the value of the fraction, the results were binned in three different bins. As we are dealing with four person meetings the average value after step 1 is 0.25 (=25% share). The features were grouped using the labels 'High' ($F'_{Pn} > 35\%$), 'Normal' ($15\% < F'_{Pn} < 35\%$), and 'Low' ($F'_{Pn} < 15\%$).

As a consequence, apart from the fact that features were now comparable between meetings, the feature values that originally had 'approximately' the same value now also ended up in the same bin. This seemed intuitively the right thing to do. Table 13 shows the value of the NOW feature ('The number of words used' per participant per meeting) before and after applying the process. If we look at the number of words used for person 2 (P2) and person 4 (P4) we see that they both end up labelled as 'High'. Although they did not speak the same amount of words, they both used more than 90000 words, which is a lot in comparison with P1 (38914) and P3 (26310), both ending up classified as 'Low'.

	NOW before preprocessing				NOW after preprocessing			
	P1	P2	P3	P4	P1	P2	P3	P4
M1	38914	93716	26310	98612	low	high	low	high
M2	33458	11602	14556	37986	high	low	low	high
M3	3496	7202	8732	2774	low	high	high	low
M4	2240	1956	4286	7642	low	low	normal	high
M5	4470	1126	9148	1974	normal	low	high	low
M6	2046	17476	1828	4058	low	high	low	high
M7	4296	6812	8258	1318	normal	high	high	low
M8	1586	13750	1786	1540	low	high	low	low

Table 13: The feature 'Number of Words' before and after preprocessing for person 1,2,3 and 4 respectively for each meeting.

Now, as the feature values were made comparable, we were almost ready to train our model. The only step left was to define the class labels determining the dominance level. For this we decided to use the same technique as for the features, labelling them also as 'High', 'Normal' and 'Low'. We calculated the shares of each of the participants by dividing the sum of the valuations of all judges for this participant by the total amount of points the judges could spend ($5 * (1+2+3+4) = 50$).

The results were then again binned using the same borders of 15 and 35 percent. Where a share was smaller than 15% the dominance level was labelled as 'High'; if the share lay between 15% and 35% the dominance level was labelled 'Normal' and where it was higher than 35 % the label 'Low' was used. This way, also the persons who received more or less similar scores ended up in the same bin.

This resulted in a data-set of 32 samples with twelve samples receiving the class label 'High', ten 'Normal' and ten 'Low'. We define our baseline performance as the share of the most frequent class label ('High') having a share of 37.5% of all labels.

4.5 Detecting dominance

We wanted to predict the dominance level of the meeting participants with the least possible features, in accordance with Occam's razor [19], trying to explain as much as possible with as little as possible. The fewer features we required, the easier it would be to eventually provide all information to the system. This way we reduced the risk of over fitting our model to the data as well. To decrease our amount of features we applied dimensionality reduction using principal component analysis.

We obtained the best performance by training a Support Vector Machine (SVM) using the two most discriminative features: NOF and NOT. These features appeared together with the NSI as being the most discriminative. Ten-fold cross validation resulted in a performance of 75%, which is much higher than our 37.5% baseline. This means, that given the number of times the meeting participants are privately addressed and given the number of times they grab the floor, our classifier is in 75 % of the cases able to correctly classify the behavior of the participants as being 'Low dominant', 'Normal dominant' or 'High(ly) dominant'. The confusion matrix is shown in Table 14.

	Low	Normal	High
Low	9	0	1
Normal	3	5	2
High	0	2	10

Table 14: Confusion matrix using the features NPA and NOF. The rows are showing the actual labels and the columns the labels resulting from the classifier.

From the confusion matrix it can be seen that our classifier performs better on the classes 'Low' and 'High' than on the class 'Normal'. This seems in line with our intuition that people showing more extreme behavior are easier to classify.

The 90% confidence interval for our classifier lies between a performance of 62% and 88%. This confidence interval is important due to the relatively small sample of data. The lower bound is still much higher than the 37.5% baseline. The fact that we would over fit our classifier when using all the features appeared when we trained on all the features. Ten fold cross validation resulted in that case in a performance of 50%.

Aware of the fact that our sample size is relatively small and that not all meetings follow the same format, we do think that our results suggest that it is possible to have a system analyzing the level of dominance of the meeting participants. If we look at the features used by our model, and the fact that their values should be just as informative during the meeting as after the meeting, we expect these systems not to function just after the meeting, but just as well in real time.

4.6 Transferring our knowledge

We used the information from our classifier to create a module for the Twente Meeting Browser where the dominance levels of meeting participants is calculated in real-time and graphically visualized in a graph.

As revealed by the SVM attribute evaluator, the features NOT, NOF and NSI were the most important. We used these to calculate a measure which we called *the dominance level*. This value is calculated as follows: We keep track of four bins, one for each participant and add points as the meeting proceeds. We decided to add one one point to the score of a participant, and in case he or she takes a turn, and add one extra point, if this turn either was obtained after a silence longer than two seconds, or by interrupting the previous speaker. At the end of the meeting, the resulting levels should match the hierarchies used to train the classifier. At the moment of writing we cannot yet confirm this for all meeting. Preliminary results however indicate that this indeed will happen. A visualization of the AMI-FOB6 meeting in the Twente Meeting Browser, including the relative dominance values is shown in figure 5. For more information about the Virtual Meeting Room, the reader is referred to Reidsma et al. [92].

4.7 Conclusions and future work

We showed that in the future systems might be extended with modules able to determine the relative dominance level of individual meeting participants. We were able to reach an accuracy of 75%. This classification appeared mainly dependent on the number of floorgrabs and the number of turns someone took. Also the number of times a

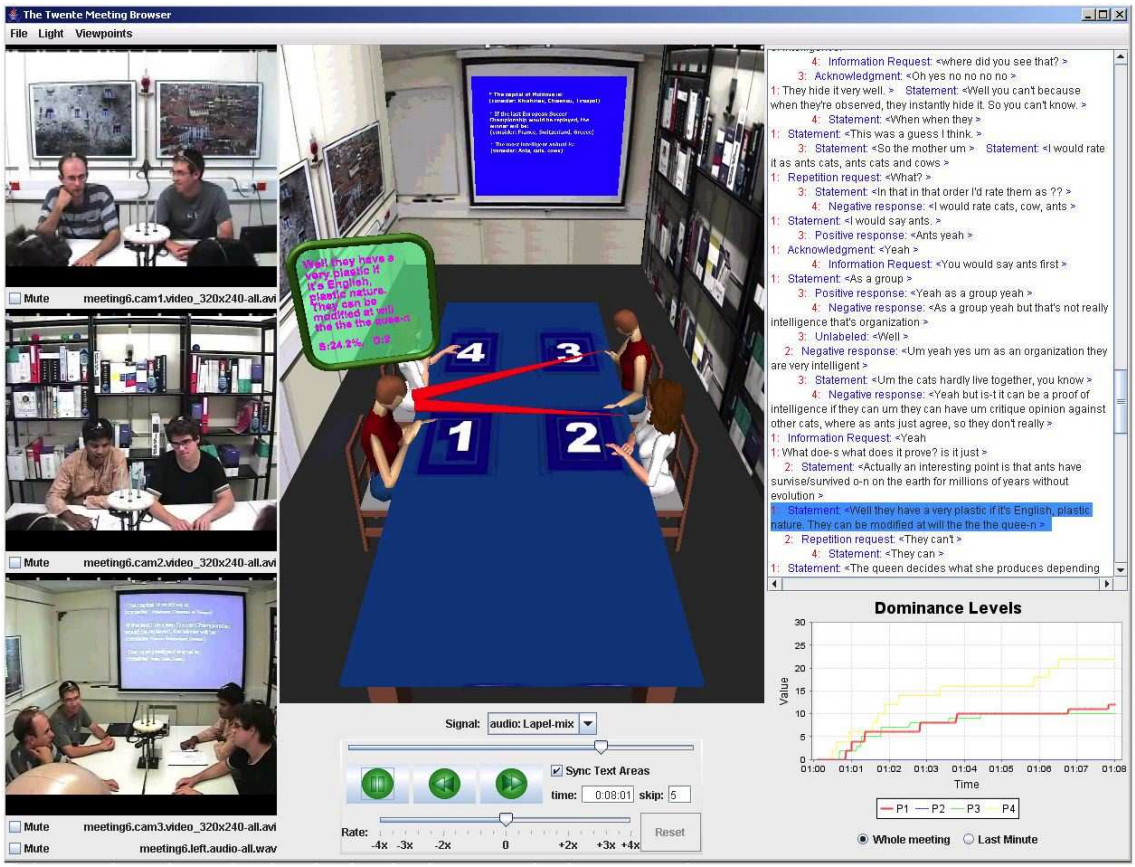

Figure 5: A view at the Twente Meeting Browser, including a dominance graph

person is privately addressed seems a good indicator in combination with the number of times the floor is grabbed by that person. As all the features are made relative to the total value of all participants, one should be able to apply the model both during as well as after the meeting

Possible directions for opportunities to improve our model could be to extend the feature set with more semantically oriented features, such as 'Who is using the strongest language?', or 'Who gets most suggestions accepted?'. Although these features seem very intuitive and might increase the performance, one does have to realize that being able to measure these, costly and complex inferencing systems have to be developed.

Another possible thing to look at is to use more samples, this will be more expensive on one side, but also decreases the confidence interval, further increasing the reliability of the performance on the other side.

Typical applications of systems that track the dominance levels of participants are other systems using the dominance information in order to inform the meeting participants or a meeting chairman about this. With this information a chairman could alter his style of leadership in order to increase the meeting productivity. Combined with other information, recommender systems could be created that directly suggest how to change the leadership style. The next thing one could think of is a virtual chairman as mentioned in Rienks et al. [97] which is able to lead a meeting all by itself, giving turns, keeping track of a time-line and most important: keeping the meeting as effective and efficient as possible.

5 Topics

Further work on hierarchical topic detection by Trieschnigg is reported in [112].

5.1 Introduction

Text segmentation, i.e., determining the points at which the topic changes in a stream of text, plays an important role in applications such as topic detection and tracking, summarization, automatic genre detection and information retrieval and extraction [87]. In recent work, researchers have applied these techniques to corpora such as newswire feeds and transcripts of radio broadcasts or spoken dialogues, in order to facilitate browsing, information retrieval, and topic detection.[7, 118, 102, 35, 18, 30]

In this report, we focus on segmentation of multiparty dialogues, in particular recordings of small group meetings as in the AMI corpus. We compare models based solely on lexical information, which are common in approaches to automatic segmentation of text, with models that combine lexical and conversational features. Because tasks as diverse as browsing, on the one hand, and summarization, on the other, require different levels of granularity of segmentation, we also explore the performance of our models for both predicting all subtopic segments and predicting only top-level segments.

In addition, because we do not wish to make the assumption that high quality transcripts of meeting records, such as those produced by human transcribers, will be commonly available, we require algorithms that operate directly on automatic speech recognition (ASR) output. Compared to read speech and two-party dialogue, multiparty dialogues typically exhibit a considerably higher word error rate (WER) [77]. Experience with segmentation of broadcast news has shown that using ASR output degrades the performance of topic segmentation models [118, 102, 18]. Therefore, it is important to understand the effect on the accuracy of the different probabilistic models we have developed for segmenting meetings.

This report is divided into 6 sections. In Section 2, we discuss previous work and its relation to our work. Section 3 describes two implemented models for automatically predicting segment boundaries for both topics and subtopics, as well as our evaluation procedure. In Section 4, we investigate how machine learning techniques can be used to cope with the highly skewed class distribution inherent in the topic organization of multiparty dialogues. In Section 5, we report the experimental results of evaluating the two implemented models on human transcripts and ASR output. In Section 6, we summarize the findings and analyze possible causes for the performance degradation. In Section 7, we briefly conclude and describe areas for future work.

5.2 Previous Work

Much of the prior research on segmentation of spoken "documents" uses approaches that were developed for text segmentation, and that are based solely on textual cues. These include algorithms based on lexical cohesion [43, 107], as well as models using annotated features (e.g., cue phrases, part-of-speech tags, coreference relations) that have been determined to correlate with segment boundaries [44, 15]. Blei et al. [18] and van Mulbregt et al. [118] use topic language models and variants of the hidden Markov model (HMM) method to identify topic segments. In fact, recent systems achieve good results for predicting topic boundaries when trained and tested on human transcriptions. For example, Stokes et al. [107] report an error rate (Wd) of 0.253 on segmenting broadcast news stories; Galley et al. [43] report an error rate (Pk) of 0.264 (when the number of segments is given) and 0.319 (when the number of segments unknown) for the task of predicting top-level segments in meetings. [11]

Although recordings of dialogue lack the distinct segmentation cues commonly found in text (e.g., headings, paragraph breaks, and other typographic cues), they contain acoustic and conversation-based features that may be of use for automatic segmentation. Acoustic information includes prosodic features [102] and speaker-specific pitch activity [9]. Conversation-based features include those obtained statistically, such as silence, overlap rate,

[11] For the definition of Pk and Wd, please refer to section 3.5.

speaker activity change [43] and cross-speaker linking information such as adjacency pairs [128], as well as those obtained empirically, such as control shift [122]. Because many of these features can be expected to be complimentary, researchers have explored approaches to select and combine features into an integrated model. For two-party dialogue, Shriberg et al. [102] have shown that combining prosodic information and lexical cues yields better results than using either alone. With respect to spontaneous multiparty dialogue, Galley et al. [43] have shown that a model integrating lexical and conversation-based features outperforms the model using only lexical features.

However, as noted above, we expect the high WER of ASR output to degrade performance of segmentation models that were developed on either human or ASR transcriptions. In particular, we expect that incorrectly recognized words will impair the robustness of text-based approaches and extraction of conversation-based discourse cues. However, no prior study has reported directly on the extent of this degradation on the performance of automatic topic segmentation in spontaneous multiparty dialogue. Past research on topic segmentation in broadcast news using ASR transcription has shown performance degradation from 5% to 38% using different evaluation metrics [118, 102, 18]. In this report, we extend prior work by providing quantitative results of applying our segmentation models to both the topic prediction and subtopic prediction tasks, and also report the results of the effect of using ASR output on models using text-based approaches and models integrating text-based and conversation-based features. For practical reasons, we leave implementation of models that integrate acoustic features to future work.

5.3 Method

5.3.1 Data

In this study, we used the ICSI meeting corpus (LDC2004S02) as a test bed for our analysis and experiments. Seventy-five natural meetings of ICSI research groups were recorded using close-talking far field head-mounted microphones and four desktop PZM microphones. The corpus includes human transcriptions of all meetings. We used ASR transcriptions of all 75 meetings which were produced by Anonymous (2005), with an average WER of roughly 30%.

Three human annotators at our site used a tailored tool to perform topic segmentation in which they could choose to decompose a topic into subtopics, with at most three levels in the resulting hierarchy. Annotators were asked to provide a free text label for each topic segment; they were encouraged to use keywords drawn from the transcription in these labels, and we provided some standard labels for non-content topics, such as ”opening” and ”chitchat”, to impose consistency,

To establish reliability of our annotation procedure, we calculated kappa statistics between the annotations of each pair of coders. Our analysis indicates human annotators achieve $\kappa = 0.79$ agreement on top-level segment boundaries and $\kappa = 0.73$ agreement on all subtopic boundaries. The level of agreement confirms good replicability of the annotation procedure.

5.3.2 Fine-grained and coarse-grained topic organization

We characterize a dialogue as a sequence of topical segments that may be further divided into subtopic segments. For example, the 60 minute meeting Bed003, whose theme is the planning of a research project on automatic speech recognition can be described by 4 major topics, from “opening” to “general discourse features for higher layers” to “how to proceed” to “closing”. Depending on the complexity, each topic can be further divided into a number of subtopics. For example, “how to proceed” can be subdivided to 4 subtopic segments, “segmenting off regions of features”, “ad-hoc probabilities”, “data collection” and “experimental setup”. For our initial experiments with automatic segmentation at different levels of granularity, we flattened the subtopic structure and consider only two levels of segmentation–top-level topics and all subtopics.

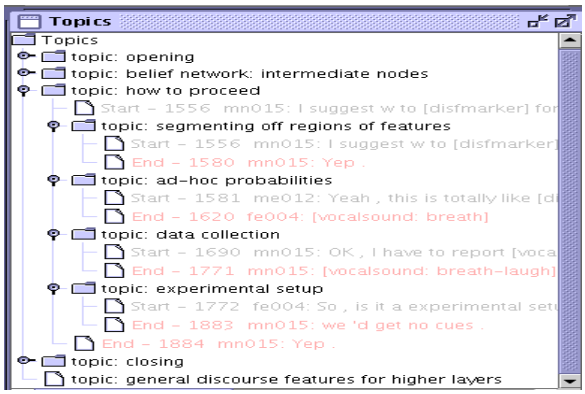

Figure 6: *The hierarchical topic structure of contents of an example meeting.*

5.3.3 Probabilistic models

To investigate the impact of ASR errors on the selection of features and the choice of segmentation models, we compare segmentation models using different types of features: (1) a model using solely lexical cohesion information, and (2) combined models integrating text-based and conversation-based features.

5.3.4 Lexical Modeling

In this study, we re-implemented Galley et al.'s [43] LCSeg algorithm, a variant of TextTiling [54]. LCSeg hypothesizes that major topic shifts are likely to occur where strong term repetitions start and end. The algorithm works with two adjacent analysis windows, each of fixed size, in our case 11 utterances. For each utterance boundary, we determine a lexical cohesion score by computing the cosine similarity at the transition between the two windows. Low similarity indicates low lexical cohesion, and a sharp change of lexical cohesion score indicates a high probability of an actual topic boundary. The principal difference between LCSeg and TextTiling is that LCSeg measures similarity in terms of lexical chains (i.e., term repetitions), whereas TextTiling computes similarity using word counts.

The first step of lexical modeling is typically to normalize the data by tokenizing, removing speaker identification information, lowering all upper case, removing function words, and stemming. However, initial results show that removing function words and stemming can impair the performance when using a lexical model for predicting top-level topics, especially on ASR output. Therefore, for the lexical model used in our experiments, we do not remove function words and do not perform stemming.

5.3.5 Integrating lexical and conversation-based features

As discussed in Section 2, prior research has shown that combining lexical information with conversation-based features outperforms a model using lexical features alone. To determine whether this is also the case when we consider the problem of predicting all subtopic boundaries and when ASR transcriptions are used, we also implemented feature-based models to learn the best indicators of topic boundaries using decision trees (c4.5), support vector machines and maximum entropy. To incorporate multiple features in the combined models, we consider topic segmentation as a binary classification task. Given a feature set and a training set with each potential topic boundary[12] labeled as either positive (POS) or negative (NEG), the classifier learns the posterior probabilities. The trained model is then used to predict whether each unseen example in the test set belongs to the class POS or NEG. In Section 5, we analyze the results of the best performing model, which is the one obtained using decision trees.

[12] In this study, the end of each speaker turn is a potential segment boundary. If there is a pause of more than 1 second within a single speaker turn, the turn is divided at the beginning of the pause creating a potential segment boundary.

For this study, we used features and the optimal window size that have been proven to perform best in prior work [43]. In particular, the results reported in this study were obtained using the following features: (1) lexical features: the raw lexical cohesion score and probability of topic shift indicated by the change in lexical cohesion score, and (2) conversation-based features: the number of cue phrases in the analysis windows preceding and following the potential boundary, similarity of speaker activity (measured as a change in probability distribution of number of words spoken by each speaker) preceding and following each potential boundary, speaker overlap rate following each potential boundary, and the amount of silence between speaker turns preceding each potential boundary.

5.3.6 Evaluation

As a first step, we performed 25-fold leave-one-out cross validation on the set of 25 meetings that were used in the study performed by Galley et al. [43]. We repeated the procedure to evaluate the accuracy using the lexical and combined models on both human and ASR transcriptions. In each evaluation, we used the automatic segmentation model for two tasks: predicting all subtopic boundaries (ALL) and predicting only top-level boundaries (TOP). The results are reported in Section 5.

5.3.7 Topline and Baseline

To compute a topline for the accuracy of our automatic segmentation models, we examined the agreement of human annotators on the task of predicting top-level segments. For the 25 meetings that were used in Galley et al.'s [43] study, we have top-level topic boundaries annotated by coders at Columbia University (Col) and in our lab (UEDIN). Following Galley et al. [43], we take the majority opinion on each segment boundary from the Columbia annotators. For the UEDIN annotations, where multiple annotations exist, we choose one randomly. The topline is then computed as the Pk score comparing the Columbia majority annotation to the UEDIN annotation.

To compute a baseline, we follow Kan [60] and Hearst [54] in using Monte Carlo simulated segments. For the corpus used as training data in the experiments, the probability of a potential topic boundary being an actual one is approximately 2.2% for all subtopic segments, and 0.69% for top-level topic segments. Therefore, the Monte Carlo simulation algorithm predicts that a speaker turn is a segment boundary with these probabilities for the two different segmentation tasks. We executed the algorithm 10,000 times on each meeting and averaged the scores to form the baseline for our experiments.

5.3.8 Evaluation metrics

Because precision and recall do not fully capture the near-miss phenomenon important for judging the performance of a segmentation model, we report our results using the standard metrics of Pk and Wd. Pk [15] is the probability that two utterances drawn randomly from a document (in our case, a meeting transcript) are incorrectly identified as belonging to the same topic segment. WindowDiff (Wd) [87] calculates the error rate by moving a window across the meeting transcript counting the number of times the hypothesized and reference segment boundaries are different. Choosing Pk and Wd as our metrics allows us to compare our results directly with previous work.

5.4 Coping with an Imbalanced Class Distribution

Previous research demonstrates that probabilistic topic segmentation models can infer high-level topic organization from low level features. However, in our context of spontaneous multiparty dialogue, the lack of a macro-level segment unit, such as paragraph or story breaks, makes the task different from the segmentation of text or broadcast news. For example, for the task of segmenting expository texts, the chance of each paragraph break being a topic boundary is 39.1% [54], while in the ICSI corpus, the chance of each speaker turn being a subtopic segment boundary is just 2.2%, and is only 0.69% for top-level boundaries. This imbalance in the class distribution affects the accuracy of the models which are trained on the imbalanced data set. Therefore, to understand the full potential

of automatic segmentation of topic boundaries in multiparty dialogue, we must tackle the problem of rare class prediction.

There have been attempts to tackle the rare class prediction problem in the fields of fraud detection, network intrusion, and web mining [29]. In the field of natural language processing, this problem is also commonly encountered in text categorization, sentence boundary detection, and disfluency detection [72].

In this study, we investigated a variety of sampling approaches, suggested in Liu et al. [72], on a data set of 25 meetings to identify the most useful approach for this task. Experiments with undersampling, oversampling, boosting and bagging to re-balance the class distribution, indicated that undersampling provides the most stable improvement in accuracy. Undersampling is a technique that removes negative examples so that the model can learn more from the positive cases during the training process. We adopted a strategy similar to Zhang and Mani's [132] direct undersampling technique, which removes the N negative examples that are closest in time to positive examples. N varies as the as the desired ratio of negative to positive examples varies. We vary the ratio to provide insight into what class distribution results in the best accuracy of the classifiers.

Another approach to coping with the rare class prediction problem that does not change the natural class distribution is to gather more instances of the rare class by increasing the size of the training set. The statistics in Table 15 show that for 25 and 75 meetings, the class distribution is roughly the same, and hence by increasing the number of meetings used in the training set we increase positive instances without distorting the natural class distribution. To explore how the change of training set size impacts the performance of segmentation models, we conducted an experiment in which we incrementally increased the training set size by randomly choosing five meetings each time until all meetings were selected. We executed the process three times and averaged the scores to obtain the results in Section 5.5.

	25 Meetings		75 Meetings	
	ALL	TOP	ALL	TOP
Training set	35238 speaker turns		108440 speaker turns	
Total topics	475	149	1717	502
Percentage of positive cases	2.42%	0.71%	2.20%	0.69%

Table 15: *Statistics of the data sets used for predicting the top-level topic (TOP) and all subtopic (ALL) boundaries. The second column shows statistics for the 25 meetings used in the initial trial. The third gives statistics for all 75 meetings.*

5.5 Results

The segmentation models were trained on all 75 ICSI meeting transcripts annotated with topic segment boundaries. A total of 6 features were used as described in Section 5.3.5. Table 16 shows the performance of the lexical model and two combined models. CM1 combines the lexical and conversation-based features discussed in Section 5.3.5. CM2 uses the same features set as CM1, and we apply directed undersampling with a ratio of negative to positive cases of 1.

As expected, the results show that both the lexical model and the combined models are more accurate for predicting segment boundaries from human transcriptions than from ASR output. For the task of predicting top-level topics from human transcripts, there is little difference in performance of the lexical and combined models. However, when using ASR output, CM1, the combined model without undersampling, is considerably better than the lexical model and CM2.

For the task of predicting all subtopics, in general, we observe that the lexical model alone is competitive with the best performing combined model and achieves accuracy that is comparable with human performance for segmenting human transcripts. However, using ASR output has a more severe impact on the accuracy of the lexical

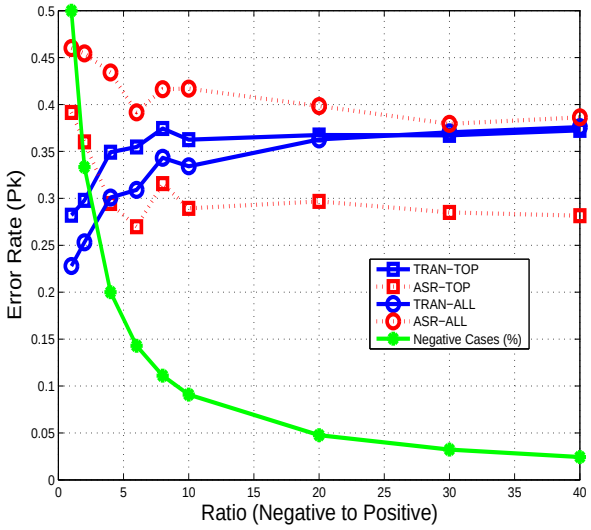

Figure 7: *Effect of undersampling on error rate*

model than on the combined models. Although none of the segmentation models we considered perform well on predicting all subtopics when using ASR output, the performance of the lexical model degrades dramatically to the baseline, while CM1 does not degrade severely.

From these results, we can conclude that when using human transcripts to predict all topic segments, the lexical model is to be preferred, but when using ASR transcription, the combined model without undersampling is most accurate. For predicting only top-level segments, there is a slight preference for CM2 when using human transcripts, but a there is a much stronger preference for using CM1 on ASR output.

Error Rate (Pk)		LM	CM1	CM2	Base line	Top line
ALL	Tran	0.19	0.36	0.23	0.47	0.18
	ASR	0.46	0.41	0.46	N/A	N/A
TOP	Tran	0.32	0.31	0.28	0.48	0.13
	ASR	0.48	0.32	0.39	N/A	N/A

Table 16: *Performance of probabilistic segmentation models in terms of error rate Pk. The topline performance for predicting ALL topics on human transcripts is obtained by comparing Columbia's top-level segments with UEDIN's all subtopic segments.*

Figure 7 shows the results of undersampling for the topic segmentation task. Note that the error rate decreases as the ratio of negative to positive examples decreases in the training set when using human transcripts. The improvement in accuracy is especially evident for predicting all subtopic segments, where there is a reduction of error rate of 37.5%, from 0.36 to 0.23. However, undersampling does not improve accuracy when using ASR transcription.

Figure 8 shows the effect of training set size on error rate for predicting top-level and all subtopic segment boundaries, with human and ASR transcriptions, and with and without undersampling. We see that increasing the size of the training set does not improve the accuracy of segment boundary prediction for any of the models. This is true regardless of whether the task is predicting all subtopics or just top-level topics, regardless of whether

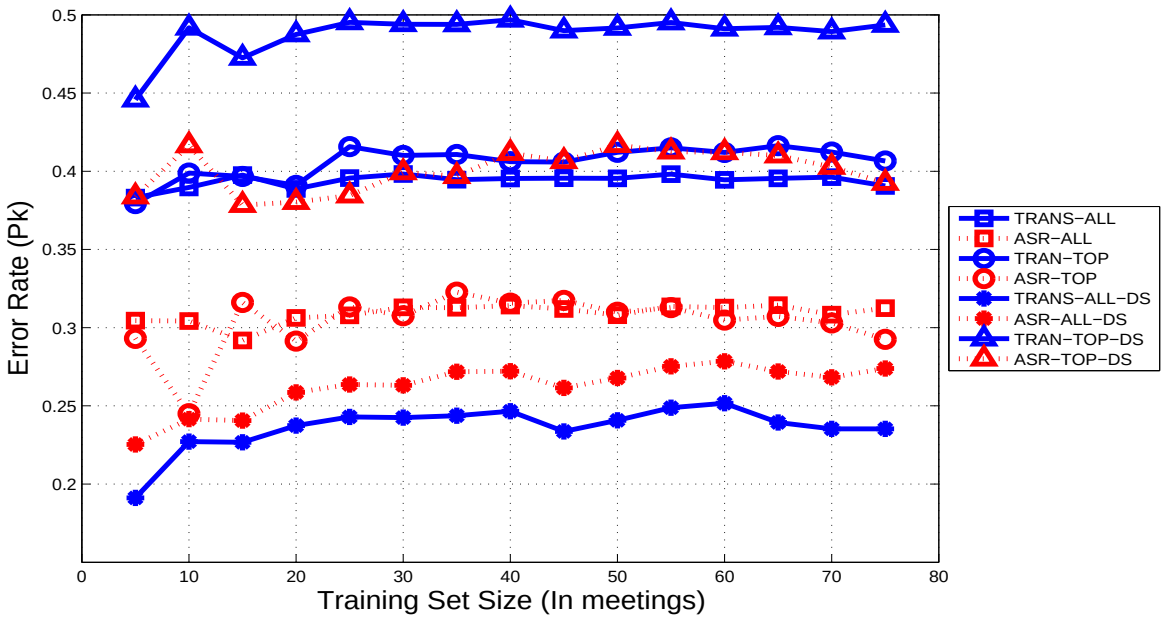

Figure 8: *Performance of the combined model (measured by error rate Pk) over the increase of the training set size.*

the input is human transcripts or ASR output, and regardless of whether undersampling is applied. That is, the accuracy level is quite stable once the training set size reaches 25 meetings.

5.6 Discussion

The purpose of modifying the class distribution of the training set is to improve the accuracy of automatic segmentation models. By using directed undersampling to reduce negative examples, we expected the trained models to be more accurate when classifying unseen data. Examination of the effect of directed undersampling for the combined models shows that (1) rebalancing the class distribution does improve the accuracy of automatic segmentation models when using human transcripts, and (2) the improvement is more evident when the ratio of negative to positive cases moves from the natural class distribution to 1. An additional advantage of directed undersampling is that it reduces training time without compromising the results, as the complexity of the combined model is a function of the total number of cases. However, directed undersampling was not effective when applied to ASR output. We hypothesize that this is because the undersampling process increases the relative importance of speech recognition errors for the negative examples.

Although we expected that including more positive examples by increasing the size of the training set would improve the accuracy of prediction, the results show that increasing training set size does not actually increase the accuracy of the trained models, regardless of whether the natural class distribution is distorted in the training set. However, the stability of accuracy level after increasing the size to 25 meetings demonstrates a possible minimum size for effective training.

5.7 Conclusions

The current study demonstrates that a lexical model alone can achieve competitive results for predicting topic segment boundaries when using human transcripts, but that a model that combines lexical and conversation-based features suffers less degradation in accuracy when using ASR output. The findings confirm that conversation-based features are more robust to incorrectly recognized words in ASR output. In order to further improve the accuracy of the combined models, we will explore the use of acoustic and other multimodal features. For example, Shriberg et al. [102] showed that combining prosodic and lexical information increases the accuracy of automatic segmentation in two-party dialogue. In addition, in the current study, we only extracted features from within the analysis windows immediately preceding and following each potential topic boundary. In future work, we will explore models that take into account features of longer range dependencies.

6 Named Entities

6.1 Task Definition

The named entity (NE) task involves identification of words or word sequences that may be classified as proper names, or as certain other classes such as monetary expressions, dates and times. This is not a straightforward problem. While 'Wednesday 1 September' is clearly a date, and 'Alan Turing' is a personal name, other strings, such as 'the day after tomorrow', 'South Yorkshire Beekeepers Association' and 'Nobel Prize' are more ambiguous. For annotation of AMI data, we essentially follow 'AMI Named Entity Guidelines'[13], which should be read as an addendum to the NIST 1999 NE recognition task definition, version 1.4[14].

The latter specification defined ten classes of named entity: three types of proper name (<location>, <person> and <organization>) three types of temporal expression (<date>, <time> and <duration>) and four types of numerical expression (<money>, <measure>, <percentage> and <cardinal>). According to this definition the following NE tags would be correct:

<date>Wednesday 1 September</date>
<person>Alan Turing</person>
the day after tomorrow
<organization>South Yorkshire Beekeepers Association</organization>
Nobel Prize

'The day after tomorrow' is not tagged as a date, since only *absolute* time or date expressions are recognised; 'Nobel' is not tagged as a personal name, since it is part of a larger construct that refers to the prize. Similarly, 'South Yorkshire' is not tagged as a location since it is part of a larger construct tagged as an organisation.

Specially for AMI data, it was decided that meeting participants and other artifacts that might be relevant to the meeting, such as furniture and recording devices, would be annotated. 'AMI Named Entity Guidelines' provides the full detail.

6.2 Annotation

The NE annotation tool has been implemented by the University of Twente, that supports the task definition described above. The annotation work is currently on going at the University of Edinburgh.

6.3 Evaluation

NE identification systems are evaluated using an unseen set of evaluation data: the hypothesised NEs are compared with those annotated in a human-generated reference transcription. In this situation there are two possible types of error: *type*, where an item is tagged as the wrong kind of entity and *extent*, where the wrong number of word tokens are tagged. For example,

<location>South Yorkshire</location> Beekeepers Association

has errors of both type and extent since the ground truth for this excerpt is

<organization>South Yorkshire Beekeepers Association</organization> .

These two error types each contribute 0.5 to the overall error count, and precision (P) and recall (R) can be calculated in the usual way.

Evaluation of spoken NE identification is more complicated than for text, since there will be speech recognition errors as well as NE identification errors (i.e., the reference tags will not apply to the same word sequence as the

[13] http://wiki.idiap.ch/ami/NamedEntities
[14] http://www.nist.gov/speech/tests/ie-er/er_99/doc/ne99_taskdef_v1_4.ps

hypothesised tags). This requires a word level alignment of the two word sequences, which may be achieved using a phonetic alignment algorithm developed for the evaluation of speech recognisers. Once an alignment is obtained, the evaluation procedure outlined above may be employed, with the addition of a third error type, *content*, caused by speech recognition errors. The same statistics (P and R) can still be used, with the three error types contributing equally to the error count.

7 Propositional Content

While dialog acts (see section 2) provide information about certain *functional* aspects of an utterance, they do not tell much about the *meaning* of these utterances and their contribution to the current discourse. For instance, both of the following excerpts from a meeting transcription might be annotated with the dialog act `question`:

- A: “How are you doing?”
- A: “What color should the power button have?”

Yet, both question differ greatly with respect to their intended meaning: the first one is a typical human-social interaction while the second is about some material aspects of a physical entity. Differentiations like this can not be delivered by the dialog act `question` alone; to encode the meaning of an utterance, we are in need of a more expressive annotation scheme to accompany the dialog act annotation.

Still, the information encoded in the dialog act is valuable for the understanding of an utterance. The information therein can be viewed as on a orthogonal axis. For instance, both of the sentences

- “Is it green?”
- “It is green.”

could be considered to carry exactly the same propositional content. The difference between them is, however, the function expressed by the corresponding dialog acts: sentence 1 is a `question`, sentence 2 a `statement`. We conclude that for a formalization of the *meaning* of discourse, dialog act annotation and propositional content annotation complement each other.

A propositional content scheme differs in principle from any other annotation scheme: while usually a scheme consists of a finite set of annotation labels one of which can be assigned to each observable annotation unit, the number of different meanings a speaker could convey with an utterance is practically unlimited. In theory, a propositional content formalism would need to have the expressibility to represent each and every meaning possible – thereby providing a complete knowledge representation of the world. Obviously, this is too ambitious a goal to be reached. On the same note, it is common for an annotation scheme to contain some sort of `unclassifiable` label because there must be a way to annotate effects that could not have been anticipated by the time the scheme was designed. The same label could also be used for effects that occur too seldom to be given their own distinct labels.

It is admissible for an annotation scheme not to cover each effect that might occur in a real meeting; still, for the case of propositional content annotation the question is legitimate whether it is possible at all to design a scheme that captures a sufficiently large percentage of possible *meanings* to be of any practical value - in an open domain application, this is at least doubtable. However, AMI hub scenario meetings which are limited to a restricted domain, the design of a remote control, give reason to expect a feasible implementation.

In order to develop a sufficient formalism for propositional content coding of natural language, we fall back on analytical philosophers classic knowledge representation mechanisms combined with modern representation formalisms.

For a long time, traditional AI has used the term ontology as coined by Gruber’s definition “An ontology is a specification of a conceptualization.” [51]. This constructivistic approach doesn’t commit itself to the representation of reality, but is restricted to the representation of particular perceptions of some part of reality and therefore can be seen as solipsitic domain ontology modeling.

In philosophy, ontology is the science of what is, of the kinds of structures of objects, properties, events, processes and relations in every area of reality. It deals with the *a priori* nature of reality and tries to provide a “definitive and exhaustive classification of entities in all spheres of being” [103]. [57] and [58] distinguish between *formal* ontologies, i.e. universal, domain-independent ontologies and *material* ontologies, i.e. domain-specific ontologies.

7.1 An Ontology for the AMI Hub Meetings

We follow the philosophical approach for several reasons. First, the philosophical underpinnings of formal ontologies ensure a framework for the creation of robust and interoperable domain ontologies and secondly, only a realistic ontology captures all aspects of communication and therefore all possible aspects of meetings.

Currently there exist only few implementations of sophisticated formal ontologies, see e.g. [52]. The most prominent of these so-called "Upper-Level Ontologies" (or "Upper Models") are the *Suggested Upper Merged Ontology* (SUMO), the *CYC upper model*, DOLCE, a *Descriptive Ontology for Linguistic and Cognitive Engineering*, the SUMO-DOLCE hybrid SmartSumo and the Component Library Ontology CLib.

All of these foundational ontologies support several basic philosophical assumptions and ontological choices at different levels of expressivity, e.g. abstract vs. concrete entities, the 3D (endurantist) vs. the 4D (perdurantist) view or a multiplicative vs. a reductionistic view and they support additional theories like mereology, topology, granularity and scale.

While SUMO and CYC are considered to be very extensive ontologies with a broad coverage and lots of mid- and domain-level ontologies, they have weaknesses regarding their axiomatisations and complexity respectively. Clib adopts some ideas from CYC and combines them with FrameNet and WordNet. Regarding their extend, SmartSumo and DOLCE are light-weight ontologies. In contrast to SmartSumo which goes without any axiomatisation and the other ontologies, DOLCE has a strong philosophical foundation. It's the only formal or upper ontology in the strict sense of Husserl's definition and can be seen as a reference module which can serve as the starting point for the development of ontologies. Since very recently, there now exists a new modularized OWL-version of DOLCE Lite Plus, enriched with experimental modules for Plans, Information Objects, Semiotics, Temporal relations, Social notions, etc.

Current ontologies are represented in a variety of languages (KIF, CycL, RDFS, KM and other proprietary formalisms); we have opted for OWL, the RDF/XML-based W3C standard for the semantic web, for three reasons. First, the OWL Ontology Language is a universal medium for the exchange of data where data can be shared and processed by automated tools as well as by humans. It's an open standard and widely used for the specification of ontologies and also there are a large number of development tools and reasoners. And last, the OWL language supports the open world assumption, which means that information that hasn't been explicitly added to a knowledge base is assumed to be "missing" information, which could be added sometime in the future.

Several material ontologies (e.g. communicative acts, meeting room, meeting, product, design, meeting, project) have been specified for the representation of the AMI hub meetings. Primarily they are based on the growing AMI corpus with currently about 715.000 words and several domain theories, e.g. theories about communication, social acting in meetings, project planning, organisations, and contributions from the AMI project, e.g. the AMI Dialog Act theory, AMI Named Entities, AMI Meeting Acts, etc.

At present the AMI ontology comprises about 2700 concepts and 600 properties, build upon the OWL version of DOLCE Lite Plus and parts of SUMO, SmartSUMO and CLib, but this may change in future due to the rapid evolution of formal ontologies. Figure 9 shows an excerpt of the AMI ontology as seen with the browsing and editing tool Protégé.

Currently we concentrate on the identification, collection and representation of the content bearing parts of the existing hub meetings, e.g. the material entities in the meeting room (whiteboard, table, chair, human, remote control, projector, microphone, etc.), the material remote control domain (remote control, button, wheel, DVD, TV, VCR, etc.), entities in the meeting domain (agenda, agenda item, participants, meeting date, meeting location, formal meeting acts, etc.) and roles (meeting manager, interface designer, product). On the basis of these domain entities we'll continue the elaboration of a generic project ontology and a meeting discourse model.

7.2 Outlook

Since DOLCE Lite Plus is a domain-independent ontology and the universal medium OWL offers the possibility to refer, share and process all kind of data, it's possible to include and map every kind of information bearing entity to some corresponding ontology entity. In the context of the AMI meeting scenario all kinds of additional

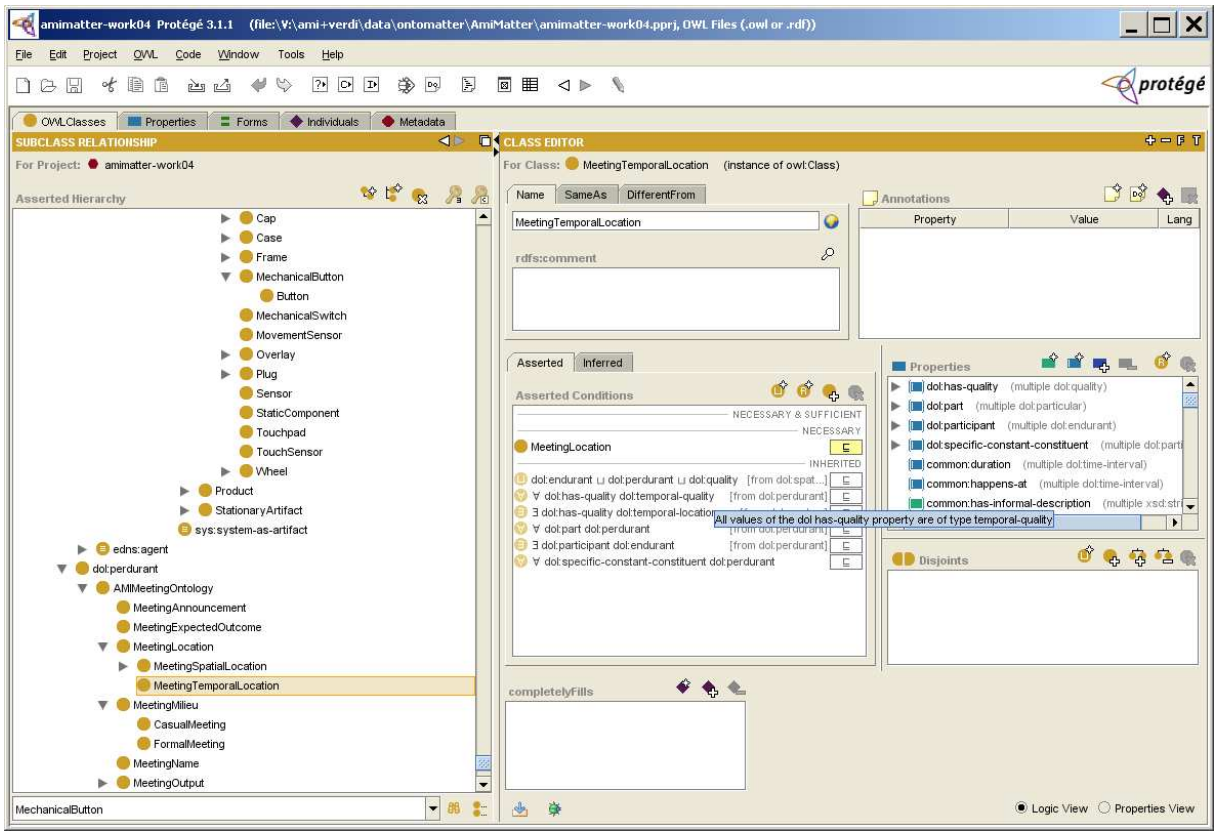

Figure 9: Screenshot of the ontology editing tool Protégé, showing an excerpt of the AMI onotology.

information entities, e.g. gestures, pictures, multimedia, powerpoint slides, handnotes, mimics etc. can be easily integrated to enrich the representation of a meeting. Another direction is the enhancement of the expressiveness of the discourse model by adding reifcation or hypostasis or the inclusion of an sophisticated interpretation theory. For example DOLCE's experimental module "Descriptions and Situations" supports the representation of complex discourses partially.

8 Structuring Meeting Data with Ontologies

In the meeting domain, ontologies can be used for the annotation of data in order to facilitate replay and navigation of meeting records, in the context of the use of a meeting browser. The taxonomical structure of the metadata is to allow users to query the data, at different levels of abstraction, from more generic, physical entities (e.g. meeting room items, participant roles, design object components), to more abstract entities (e.g. argument network and decision).

The development of the conceptual model of the meeting ontology for the AMI meeting recordings is based on the description of the design meeting scenario, in which the task is the design of a remote control. Several kinds of knowledge have been captured in the meeting ontology: generic knowledge encoded in concepts from the chosen upper model - the Component Library (CLib) - in order to ensure a common understanding of the foundational concepts of the domain. This upper model has been cut down, for our own modelling purposes; general meeting knowledge (meeting phases, items, participants, goals, actions and decisions); meeting type specific knowledge (design phases, tasks, roles, methods); domain specific knowledge (remote control physical model).

In the specific context of design meetings, the ontology-based annotated propositional content that identifies, on the one hand, initial project goals, and, on the other hand, proposals, positions, decisions, should subsequently allow not only for the recovery of the line of reasoning in the design process, but also for the measurement of the team performance, based on the meeting outcome.

Considering the fact that the argumentation structure can be expressed in different modalities, we have considered the possibility of relating verbal argumentative elements to other annotated items pertaining to individual meeting behaviour (hand/head/body gestures, postures, moods, location) for a more meaningful navigation through meeting capture.

References

[1] Niekrasz J., Purver M. and Dowding J. and Peters S. (2005) ”Ontology-based Discourse Understanding for a Persistent Meeting Assistant”, AAAI Spring Symposium, ”Persistent Assistants: Living and Working with AI”.

[2] Pallotta V. and Hatem G.(2003) ”Argumentative Segmentation and Annotation Guidelines”, Internal Report IM2.MDM.

[3] Karacapilidis N. and Evangelou C. (2005) ”Interweaveing Knowledge Management, Argumentation and decision-making in a collaborative setting: the KAD ontology model”, International Journal Knowledge and Learning, Vol. 1, Nos.1/2.

9 Argumentative Structure

Within organizations the locus of a lot of knowledge production is found in dialogue, discussion, and argument: people expressing ideas, negotiating deals, arguing viewpoints, pursuing agenda's and seeking common ground. The arena where most of this production occurs are meetings. The general visible results of meetings normally are meeting minutes, and maybe if lucky a list of action plans. Generally, a lot of energy and information that has been put into the actual outcome is never seen again. Meeting records however can also contain recordings of meetings where the whole meeting is captured by a number of camera's and microphones.

Smart meeting rooms have appeared at several institutions in order to record large corpora of meeting data aiming to eventually build models and systems able to capture the relevant content of the meeting. Once this content can be transformed into information sources, one will be able to exploit them to gain more knowledge about decision making, planning, assessment and rationale capturing [85]. This content, also known as organizational memory, can be made accessible afterwards for further study by e.g. a browser or a summarizer. For a complete overview of how technology can support meetings see Rienks et al. [97].

Lisowska [71] lists the kinds of queries people want to ask about records of meetings. Two main groups of questions are distinguished. The first deals with questions about the interaction amongst participants during a meeting. These are questions such as: *Who was in favor of the proposal from X, Where there any objections raised to the final conclusion?*, or, *Where there any other solutions debated?*. The second type deals with elements from the meeting domain itself. Examples are *How long did the meeting take?*, *Who where the attendees of the meeting?*, or *What were the issues debated, and which problems are still unresolved?*.

We are interested in finding answers to questions dealing with agreement, disagreement, discussions, decisions and arguments. We try to find an approach that is able to capture the decisions of a meeting as well as the lines of deliberated arguments. We do not want to formulate an opinion about the *contents* of the argumentation, but we do want to identify the relations and the forthcoming *structure* between the arguments. In this paper we introduce the Twente Argument Schema, which is developed in order to structure textual units by providing an annotation enabling people as well as automatic systems to find answers to questions related to the decision making process.

As design can be considered an 'ill-structured' or 'wicked problem', the approach in a collaborative problem solving process one encounters in these kinds of meetings is generally through a lot of argumentative discourse [24]. We've tried to identify the various functions of the argumentative aspects of the different contributions made by the participants and defined labels to relate these contributions towards each other. The resulting structure provides extra insight into which issues were debated and which statements were put forward. The schema contains labels for transcript fragments as well as labels for relations between these fragments. The resulting structure captures the discussions and can be aligned with models structuring arguments developed by argumentation theorists (c.f. Toulmin [111]). The examples used to illustrate the schema are mostly taken from the transcript of the AMI-FOB6 meeting, in particular the intelligence discussion which is included in Appendix I.

9.1 Argument Diagramming

The primary tool currently in use to give an account of argument structure is the argument diagram. There are many different kinds of argument diagrams. An argument diagram generally provides a map or snapshot of the overall flow and structure of the extended chain of reasoning in a given passage of discourse containing argumentation. A typical argument diagram gives a map of the overall structure of an extended argument. The diagram generally is a graph containing a set of points or vertices joined by lines or arcs. The points (nodes) are used to represent statements and conclusions of the argument, the lines (arrows) join the points together to represent steps of inference.

The first one to represent the structure of argumentation by using diagrams was Beardsley [14]. This consisted of numbered statements and arrows indicating support relationships. Coherence between various aspects of the dialogue are in this way revealed.

Argument diagrams often serve as a basis for criticism and reflection of the discussion. A related term in relation to argument diagramming is *design rationale*, which is a systematic approach to layout the reasons for and

the reasons behind decisions that led to the design of an artifact [23]. Argument diagrams can be used for various other purposes. We list them here briefly:

- Argument diagrams provide a representation leading to quicker cognitive comprehension, deeper understanding and enhances detection of weaknesses [98, 61].
- Argument diagrams aid the decision making process, as an interface for communication to maintain focus, prevent redundant information and to saves time.
  [126, 119].
- Argument diagrams keeps record and functions as organizational memory.[23, 85]
- The development of argument diagrams may teach critical thinking.[91, 116]

It is obvious that they can serve very similar functions when applied to records of meetings.

9.2 Diagramming methods

Several diagramming techniques have been developed, all with their goals in mind and their own ways to create the diagrams. We discuss three of them : Wigmore's charting method, Toulmin's model and the model developed for the EUCLID project.

Wigmore's charting method Wigmore [125] developed a graphical method for charting legal evidence and looks like the traditional diagramming methods one encounters nowadays in logic textbooks (e.g. Govier [49]). The purpose of his charting mechanism is to represent proof of facts in evidence presented on either side of a trial, to make sense of a large body of evidence. His charts depict the arguments that can be constructed from this body of evidence as well as possible sources of doubt with respect to these arguments.

In his model each arrow represents an inference or a provisional force. The nodes are the *facts* or the kinds of evidence that are put afore. Each type of evidence has its own shape. Circumstantial evidence is for example represented by a square, where as testimonial evidence is represented by a circle. Furthermore there are possibilities for including a type of relation between facts where one fact 'explains away the other', whether the evidence was offered by the defendant, or whether the fact was observed by a tribunal or judicially admitted.

The Toulmin model In the late 1950's Stephen Toulmin developed a model of where a schematic representation of the procedural form of argumentation is presented [111]. Toulmin's model is only concerned with pro argumentation and the acceptability of a claim, that is to say the role played by verbal elements in the argumentation during the justification process.

Toulmin regards an argument as a sequence of interlinked claims or reasons that between them establishes the content and force of the position for which someone is arguing. He states that an argument consists of six building blocks: A *datum* which is a fact or an observation, a *claim* related to the datum through a rule of inference which is called a *warrant*, a *qualifier* which expresses a degree of certainty of a claim, a *rebuttal* containing the allowed exceptions and a *backing*, which can be used to support a warrant.

The EUCLID Model A final model we discuss is the EUCLID model, a hypertext-like model of arguments developed under the EUCLID project. This diagramming method relies on the segmentation of a discussion into a series of claims. This model is rather simple as the resulting claims can only be related to each other by either 'support' or 'refute' links [104].

What we see is that these diagrams all have serve their own purpose and show differences in application domain or level of detail, they have one thing in common. They all have their own labels and with these labels they structure parts of discourse in a way to facilitate comprehension and point out possible flaws. As our model should be able to reveal similar structures, but not from evidence used in trials, but from meeting transcripts we are faced with other limitations. Not all argumentation will be in favor of a particular issue, neither will all the components as defined by the Toulmin model be present.

We now consider some software tools that are used for argument diagramming purposes and see what we can learn from them.

9.3 Diagramming tools

Nowadays several computer software tools are available that are able to help with the creation of an argument diagram. These Computer Supported Argument Visualization (CSAV) tools or applications are designed to assist in sorting and sense making of, information and narratives found in minutes or other forms of discourse weaving threads of coherence. Users are able to manipulate, annotate and display the structure in various ways. Although all the tools provide means for the creation of an argument diagram they all have their own underlying model or method with their own set of components from which in the end the resulting diagrams can be created. The components, or objects and relations, and the rules for combining them are referred to as the 'representational notation' [109]. We will now describe some features of these tools and look at their representational notations for defining their diagrams.

Most of these tools aim to provide a means for both students and scholars in argumentation to analyze the structure of natural argument. Araucaria [91], named after a tree, is for example such a tool. In Araucaria argument premises are to be placed below the conclusions and all nodes (propositions) and the connections between them can be labelled according to their evaluation. Another educational tool aiming to increase critical thinking is Reason!able [117], which is designed to be used in undergraduate thinking classes. The primary objects in reason!able are claims, reasons and objections. These components can be used to model argument trees. In the resulting argument trees, a 'child' is always evidence for or against a parent. Similar trees can be constructed with another software package called Athena[15] and Belvedere [108].

There are some differences between the capabilities of these tools. Araucaria is for instance able to handle argumentation schemes in a way that in case a complex of propositions is joined through an schema, the whole structure can be labelled and highlighted and has the ability to show counter arguments in a shaded box linked by an horizontal line to the proposition it counters. It is therefore also used for the creation of a collection of arguments fitting within typical argument scheme's (Katzav et al. [62]). In Athena, users are able to manually assign a relevance value to the relations and to manually evaluate the acceptability of the premises to see how much strength a parent would derive from its children. With Reason!able one is able to evaluate arguments on three different levels. The strength of the arguments (on a three level scale: no support, weak support and strong support), the degree of confidence in their truth and independent grounds for accepting or rejecting (e.g. because it was stated by an authority). The Belvedere environment allows the nodes to be labelled with labels as *Principle, Theory, Hypothesis, Claim, Data* where as in Reasonable, the nodes can be only of type *Claim*.

A somewhat different tool is Compendium [100], which was designed as a tool to support the real time mapping of discussions in meetings, collaborative modelling, and the longer term management of this information as organizational memory. Another difference with the other tools is that the resulting diagram can contain apart from arguments or conclusions also questions or issue as well as, answers or ideas that have been expressed. Furthermore decisions can explicitly be indicated as well as that references to external data sources can be included such as notes and spreadsheets.

This shows some of the tools that are used to capture argument diagrams. Also for the schema we are developing an annotation and visualization tool is being constructed. With respect to the representational notations of the tools, it appeared that the positive (support) and negative (refute) relation between arguments are included in all of the tools. Only in the Belvedere environment the relations are somewhat finer grained, examples of their relation set are *support, explain, undercut, justify, conflict*. Another observation is that in all of the tools, except compendium, the main conclusion or thesis that was debated is represented as the uppermost node.

9.4 Aspects of a dialogue

The argument diagrams discussed above visualize the structure of an argument. In many cases argument diagrams are constructed to analyze an argument that has been expounded in a text or that has been expressed through a dialogue. In this case, it is even possible that statements may be put into the diagram that were not expressed explicitly in the text. The purpose of the schema that we present in the next section is to annotate the statements

[15] www.athenasoft.org

from a text or the utterances in a dialogue with labels that indicate their argumentative function in the discourse or the argumentative relation that holds between them. In this sense, the schema attempts to capture information closely related to the kind of relations found in argument diagrams, but is in its nature closer to a dialogue act scheme or a scheme such as that stemming from Rhetorical Structure Theory.

Rhetorical structure theory from Mann and Thompson [74] provides an inventory of relations that hold between the sentences (roughly speaking) in a text that account for one aspect of coherence: what has a sentence to do with the preceding or the next sentence. The list of relations posited is open-ended. The set of relations is meant to be general, though in specific genres of texts some relations are more likely to turn up than others. Some of the relations proposed in RST are: evidence, background, elaboration, contrast, condition, motivation, concession, restatement. Some of these, such as evidence and concessions, will typically occur in argumentative discourse.

In the original set-up by Mann and Thompson [74] rhetorical relations are not considered to be speech acts. However, it is clear that they are not completely unrelated. Each of the relations could correspond to or constitute a speech act: provide evidence, give background information, elaborate, contrast, make a conditional statement, motivate, concede, restate. Asher and Lascarides [10], using rhetorical relations to account for a range of semantic processes in language, therefore consider rhetorical relations as speech acts that are relational.

For establishing the kinds of speech acts we want to use to mark the argumentative function of utterances, we have to look at the kinds of dialogues or texts that we want to consider. We are especially interested in dialogues where participants discuss the pro's and cons of certain solutions to a problem, providing arguments in favor or against the various solutions and raising new problems. This is not completely unlike the discussions that are modelled in the IBIS system. The IBIS model [66] is an approach to fit argumentation in a model in terms of issues and their alternatives that have been proposed and accepted by the participants. (Note that IBIS is not a graphical diagramming model) It is based on the principle that the design process for a complex problem is a conversation between the participants who each have their own area of expertise. In the process the problem is also called the topic. Within this topic, speakers bring up issues. Whenever speakers have an opinion towards an issue, they can assume a position to state how they look at the issue. To defend their opinion towards the issue they can construct arguments until the issue is settled. In this process the participants give their opinion and judgement about the topic and thus create a more structured look of the topic and its possible solution [32].

Important conversational moves in this kind of dialog are: raising problems, putting forward assertions (solutions), retracting assertions, and putting forward arguments in favor or against a solution. An assertion expresses a proposition and a form of speech indicating whether the assertor is committing to a specific position in a strong or a weak way. The schema that we present in detail in Section 9.6 accounts for the basic elements of these kinds of moves. It distinguishes acts in which issues are raised (questions put forward) and statements for a position that are made. It allows one to indicate whether a statement is strong or weak. Whether statements agree or disagree with each other can be marked in the relations. In many cases statements are not simply in favor or against but variations of each other: restatements, specializations or generalizations. This is something we account for as well in our schema. Before we present some further details, we will discuss some general issues that we took into consideration.

9.5 Defining our own diagramming model

As we intended to use an external graphical representation of argumentation, we had to decide on the representational notation that we could use. According to Bruggen [22] the most important question that needs to be answered is *what* the representational notation of the external representation must contain before one starts defining this notation.

Our representation should visualize the structure of our design meeting discussions containing the contributions from the meeting transcripts in a crisp and coherent way, such that answers to questions asked about the meeting either follow directly from the schema or can be derived in a straight and easy manner.

Walton and Reed [124] describe five what they call 'desiderata' for a theory of argument schemes. Although they regard argument schemes as form of an argument (structures of inference) representing common types of

argumentation, the desiderata are also relevant for models describing the components and the relations these components in order to constitute an argumentation diagram and thus relevant for our purpose.

The desiderata are:

1. Rich and sufficiently exhaustive to cover a large proportion of naturally occurring argument.
2. Simple, so that it can be thought in the classroom, and applied by students.
3. Fine-grained, so that it can be useful employed both as normative and evaluative system.
4. Rigorous, and fully specified, so that it might be represented in a computational language.
5. Clear, so that it can be integrated with the traditional diagramming techniques of logic textbooks.

These desiderata also hold for our schema.

The decision making process occurring in our design meetings can be decomposed into several sub processes, with multiple levels of detail. An example is the nine-step model proposed by Schwartz [99] which mentions the following phases: the problem definition, the criteria definition to evaluate the solutions, identify the root causes, generate solutions, evaluate solutions, select the best solution, develop an action plan, implement the action plan and evaluate the the outcomes and the process. A similar decomposition is presented by Briggs and Vreede [21] who identify structures such as, diverge, converge, organize, elaborate, abstract and evaluate. So as we want to capture the decision process of a meeting our model should somehow be able to incorporate these relevant aspects.

With respect to all diagramming models we studied, they generally start with, or work towards a final 'conclusion'. This does not suit our purpose as it could happen that in our domain of meeting discussions. there might be no conclusion at all (e.g. due to time constraints). What we would like to do is to capture contributions, or parts of contributions in the nodes of the diagram that is to be developed. Also the support and object relations with respect to issues debated seem to be appropriate for our use.

The approach that we took was a so called 'goal driven design' approach. Based on the literature on argumentation theories and argument diagramming, we started by creating argument diagrams on a small corpus. In several rounds we tried to reach a consensus on how to label a meeting. When required, the representational notation was refined. The whole process was repeated until agreement was reached on the labels for the components. The next section describes the resulting schema and relates it to components of the other models described before as well as to the structural components inherent to conversations.

9.6 The Twente Argument Schema

The Twente Argument Schema is a Schema that can be used to create argument diagrams from meeting transcripts. Following most of the diagrams studied, application results in a tree structure with labelled nodes and edges. The nodes of the tree contain parts of, or even complete speaker turns. The content of the nodes correspond in granularity to the size of dialogue acts. The edge define the type of the relation between the nodes.

9.6.1 The Nodes

As Newman and Marshall [80] describe, if one is willing to make a decomposition of large and complex spaces, a separation of issues is required that group arguments with respect to a particular topic they address. (c.f. a meeting agenda). In the IBIS model issues are represented as questions [66]. This is due to the fact that issues can be seen as an utterances with a direct request for a response, in the same way as a question is generally followed by an answer.

Fundamental questions with respect to conversational moves are *yes-no questions* and *why questions* [65]. A Yes-No question admits only two kinds of answers, being it either supportive, or negative. A yes no question rules out the *option* 'I don't know' expressing uncertainty. Both types of questions are so called choice questions where the set of possible options to answer is limited to a defined set of choices. Another type of question one could ask

is an *open question*, this question can be answered in any way without the limitation of a predefined set of choices, for the progress of the dialogue, the only restriction is that the answer should somehow be related and relevant to the question [50]. In our Schema we defined three different labels for our nodes to represent the issues: The *'Open issue'*, the *'A/B issue'* and the *'Yes/No issue'*. As a response to the issues, participants can take positions with respect to the possible set of options relevant to the issue. These positions are generally conveyed through the assertion of *statements*. The content of a statement always contain a proposition in which a certain property or quality is ascribed to a person or thing. A proposition can be a description of facts or events, a prediction, a judgement, or an advice (Van Eemeren et al. [115]).

Statements can vary in the degree of force and scope. It can happen that meeting participants make remarks that indicate that they are not sure of what they say is actually true. Toulmin [111] uses in his model a qualifier to say something about the force of what he calls 'claim'. When this qualifier is introduced, it is possible that the assertion is made with less force. As Eemeren [40] points out that the force of an argument can also be derived from lexical cues. To be able to represent this we introduce the label *'weak statement'*.

So, the nodes in our tree consist of issues and statements. Where statements can be either weak or strong and issues are distinguished in whether they are open, yes/no or present several alternatives.

9.6.2 The Relations

Relations can only exist between nodes. For this we have defined a number of relations that can exist between the labelled nodes. When engaged in a discussion or debate, the elimination of misunderstandings is a prerequisite in order understand each other and hence to proceed [79]. Participants in a discussion, according to Neass, eliminate misunderstandings by clarifying, or specifying their statements. These moves can e.g. be observed in the criteria definition phase, of the decision making process.

If one clarifies a statement, the new contribution sheds a different light on the same content to increase comprehension by the other party. As this occurs regularly in the discussions examined we introduced the *'Clarification'* relation label. It is to be noted that a clarification contribution can also be made by a different person than the person making the initial contribution. An example of a clarification relation occurs between the following two contributions in our example 'Ants are the most intelligent animals' and the proceeding contribution of the same speaker shows why this is the case 'Ants can build big structures'. The second contribution here is used to clarify the first one by explaining why the speaker thinks that what was said by his first contribution is true.

A specification occurs in situations where a question is asked by one of the speakers and someone else asks a question which specializes the first question resulting in a possible solution space with more constraints. The contribution 'Which animal is the most intelligent?' can be specialized with the following proceeding contribution 'Is an ant or a cow the most intelligent animal?' which again can be specialized if one for instance asks 'Are ants the most intelligent animal?'. The other way around is however also possible. If one is not able to find a solution for the specific problem, one could enlarge the solution space through generalization. For these occasions we introduce the labels *'Specialization'* and *'Generalization'* Both labels can for instance be applied when a particular issue generalizes or specializes another issue.

Whenever the issue is defined, an exchange of ideas about the possible answers or possible solution naturally occurs in the decision making process. Whenever a statement is made as a response to an open-issue or an A/B-issue it might reveal something about the position of participant in the solution space. In general he provides an *'Option'* to settle the issue at hand. For example when a speaker asks 'Which animal is the most intelligent?' and the response from someone else is 'I think it's an ant' the option relation is to be applied. The opposite of the option relation is the *'Option-exclusion'* relation, and it is to be used whenever a contribution excludes a single option from the solution space.

For a yes/no-issue the contributions that can be made are not related to enlarge or to reduce the solution space, but to reveal one's opinion to the particular solution or option at hand. In a conversation people can have a positive, negative or neutral stance regarding statements or Y/N-issues. For this purpose the labels *'Positive'*, *'Negative'* and *'Uncertain'* are introduced. With the aim to reveal whether contributions from participants are either supportive,

objective, or unclear. We see that the positive and negative label are used in many of the models described in section 9.2 and 9.3.

The positive relation for example can exist between a yes/no-issue and a statement that is a positive reaction to the issue or between two statements agreeing with each other. When one speaker states that cows can be eliminated as being the most intelligent animals and the response from another participant is that cow's don't look very intelligent, then the relation is positive. The negative relation is logically the opposite of the positive relation. It is to be applied in situations where speakers disagree with each other or when they provide a conflicting statement as a response to a previous statement or a negative response to a Yes/No-issue. In case it is not clear whether a contribution is positive or negative, but that there exists some doubt on the truth value of what the first speaker said, one should use the uncertain relation. From experience with the annotations it appears that in most cases it can easily be seen by the annotator whether the remark is mostly agreeing or mostly showing doubt.

The final relation of our set is to be applied when the content of a particular contribution is required to be able to figure out whether another contribution can be true or not. We named this the *Subject to* relation, which is somehow related to the concession relation in Toulmin's model. It is to be applied for example in the situation where someone states that 'If you leave something in the kitchen, you're less likely to find a cow' and the response is 'That depends if the cow is very hungry'. So the second contribution creates a prerequisite that has to be known before the first contribution can be evaluated. If the cow is very hungry the support could be either positive or negative. The uncertain label is not to be applied it this case, as the stance of the person in question is clear once the prerequisite is filled in. The uncertain label is merely to be used when a issue is preceded by a request for specialization or clarification.

9.7 Preserving the conversational flow

As we are working on transcripts, it is best for our model to be constructed sequentially in order to follow the line of the discussion. To preserve the order of the discussion in the model we decided that, when applying the schema, the algorithm or annotator should follow a depth first search algorithm [34] when extending it. This means that in principle every next contribution becomes a child of the previous contribution, unless the current contribution relates stronger to the parent of the previous contribution. This way the resulting tree structure can still be read synchronously.

9.8 Freedom of the annotator

One of the drawbacks of argument diagramming that is often mentioned is that there is no correct diagram. Walton [123] for instance showed that various different argument diagrams can be instantiated by one single text. Although this is true, an interesting point here is the analogy that can be drawn between RST and Argument Diagramming. As Reed and Rowe [91] point out that Mann and Thompson suggest that the analyst should make *plausibility judgements* rather than absolute analytical decisions, it is implicated that there can be more than one reasonable analysis. This also goes for argument diagramming, where the evaluator is free to interpret and to create that diagram that he considers the most appropriate according to his or her perception. As long as the schema is applied correctly, its purpose anyhow will be apparent. An example of a transcript can be found in Appendix A and the resulting diagram can be found in Figure 10.

9.9 Conclusions and Future Work

We have developed a method to capture argumentative aspects of meeting discussions in a way that an argument diagram can be created that shows how the discussion evolved, how the contributions of the participants relate, which issues were debated and which possible solutions were evaluated. The resulting argument maps are a valuable resource capturing organizational memory, that can aid querying systems and can be directly used in meeting browsers.

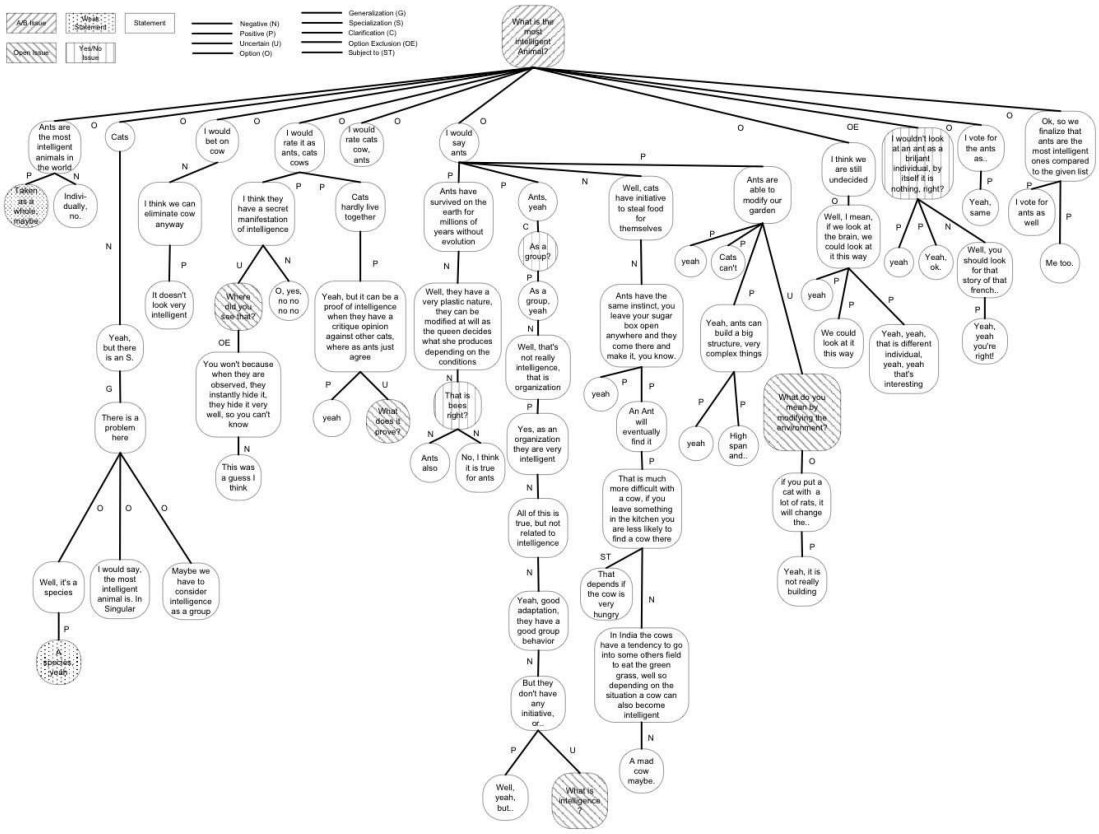

Figure 10: An argument tree of the intelligence discussion

Currently we are developing a tool to help make annotations of meeting transcripts. This way we want to construct a corpus from which models are to be trained that hopefully one day will be able to derive the structures of the argumentation themselves. When the first annotations are there, reliability analysis amongst annotators will be investigated as well as a study into automatic recognition of both labels and relations are to be conducted.

10 Chunking

10.1 Introduction

Steven Abney pioneered the idea of *parsing by chunks* supported by psychological evidence of human parser [1], where chunks are taken to be some non-recursive cores of *major* phrases. He also tried partial parsing of unrestricted text with finite-state cascades [2] in a knowledge-intensive way.

The problem of chunking is further reformulated as a task similar to POS tagging [89], i.e., by adopting a tag set of $\{$B, I, O$\}$ combined with chunk type of XP for those non-overlapping chunks, where:

B: initial word of a chunk

I: non-initial word of a chunk

O: word outside of any chunk

Therefore many learning approaches to POS tagging become directly available for chunking (see, e.g., [83, 84]).

Syntactic chunking (partial parsing) of unrestricted written text have become a relatively well-defined and well-studied task since the introduction of CoNLL 2000 shared task [110]. But the chunking of spontaneous spoken langauge has received less attention (except [84]) than that of written language though spoken language is also suitable (if not more) for such kind of shallow processing. The successful chunking of AMI meetings would serve the meeting browser in several ways, direct (e.g., to find some meaningful unit larger than words) or indirect (e.g., to extract chunk features for further analysis like segmentation, dialogue act tagging, summarization, etc).

In this section we will, on the one hand, try three different classifiers (based separately on maximum entropy/MXPOST, support vector machines/SVMs, and conditional random fields/CRFs) on Penn treebank Wall Street Journal (WSJ) and switchboard (SWBD) to show state-of-the-art performances of chunking. On the other hand, we will test AMI meetings with those chunkers to show the effect of the difference training data on chunking performance. And therefore we propose to apply semi-supervised learning to tackle the annotation problem.

10.2 Data and classifiers

The Penn treebank data used here includes WSJ sections 15-18 as training data (wsj.train), and section 20 as test data (wsj.test); SWBD sections 2 and 3 as training data (swbd.train), sections 4 as test data (swbd.test); AMI meeting IS1008b as training data, IS1008a as test data. Penn treebank is converted from trees to chunks using the script for CoNLL task. The evaluation script is the same as CoNLL. The performance is reported in $F_{\beta=1}$ score (%) unless indicated. AMI meetings are manually chunked in a similar manner to CoNLL task as described in [110].

The classifiers used are MXPOST [90][16], YAMCHA[17], and CRF++[18].

[16] Available from `ftp://ftp.cis.upenn.edu/pub/adwait/jmx/jmx.tar.gz`.
[17] Available from `http://chasen.org/~taku/software/yamcha/`.
[18] Available from `http://chasen.org/~taku/software/CRF++/`.

10.3 Experiments and Results[19]

Training on WSJ

	mxpost	svm	crf
wsj.train	92.92	99.97	99.32
wsj.test	88.35	87.85	88.55
swbd.test	71.06	70.62	71.49
IS1008a	60.91	57.57	62.01
IS1008b	61.97	58.94	60.91

Training on SWBD

	mxpost	svm	crf
wsj.test	76.70	75.39	77.78
swbd.train	92.34	99.61	97.28
swbd.test	89.93	90.96	91.82
IS1008a	74.39	70.66	72.97
IS1008b	74.20	71.48	71.91

Training on AMI IS1008b with MXPOST

test on IS1008a

	Precision	Recall	$F_{\beta=1}$
ADJP	30.30	30.30	30.30
ADVP	61.06	55.65	58.23
CONJP	93.55	91.58	92.55
INTJ	95.00	93.66	94.33
NP	81.12	86.19	83.58
PP	81.36	85.71	83.48
SBAR	63.64	63.64	63.64
VP	75.33	77.57	76.43
Overall	81.07	82.93	81.99

test on IS1008b

	Precision	Recall	$F_{\beta=1}$
ADJP	78.72	65.49	71.50
ADVP	78.77	75.91	77.31
CONJP	91.67	93.77	92.71
INTJ	96.23	97.28	96.75
NP	87.99	91.53	89.73
PP	94.08	91.53	92.78
SBAR	92.00	85.19	88.46
VP	84.67	85.32	85.00
Overall	88.32	89.07	88.69

[19] Please note: all the experiments here did not make use of POS information, simply to make things simpler. Therefore, the results can not be compared directly with those reported in most of the chunking papers, e.g., [110], where POS is used. By the way, the sota performance for chunking is around 94% in $F_{\beta} = 1$.

10.4 Discussion

From the above experiments, we come to the following conclusions:

- The sota chunking (without POS information) performance ($F_{\beta=1}$) on annotated Penn treebank data is about 87.85 - 91.82 %.
- Of all the chunkers trained on Penn treebank data, the best performance in chunking AMI meetings is from the chunker trained with MXPOST, which is also the most computationally efficient. So if we have to find a best chunker trained on Penn treebank data with any classifier, then we need to choose SWBD data and MXPOST.
- From further experiments on AMI data, training data of the same genre or domain is the most informative. But for AMI meeting chunking, we don't have any chunk-annotated data for training and there won't be any large-scale annotation. So, we will need to employ some kind of semi-supervised learning approach. Actually, annotation of a small data set is ongoing. Once it's finished, the data will be used as seed data to bootstrap an AMI meeting chunker with sota performance.

Acknowledgemet

Many thanks to Adwait Ratnaparkhi and Taku Kudo for their freely providing such wonderful toolkits.

11 Meeting Group Action Segmentation and Recognition

In this section we address the problem of recognising sequences of human interaction patterns in meetings, with the goal of structuring them in semantic terms ([3]). The aim is to discover repetitive patterns into natural group interactions and associate them with a lexicon of meeting actions or phases (such as discussions, monologues, and presentations). The detected sequence of meeting actions will provide a relevant summary of the meeting structure. The investigated patterns are inherently group-based (involving multiple simultaneous participants), and multimodal (as captured by cameras and microphones).

Starting from a common lexicon of meeting actions (section 11.1) and sharing the same meeting data-set (section 11.2), each site (TUM, IDIAP and UEDIN) has selected a specific feature set (section 11.3) and proposed relevant models (section 11.4). Then a common evaluation metric (section 11.5) has been adopted in order to compare several experimental setups (section 11.6).

11.1 Action Lexicon

Two sets of meeting actions have been defined. The first set (lexicon 1, defined in [75]) includes eight meeting actions, like discussion, monologue, or presentation. The monologue action is further distinguished by the person actually holding the monologue (e.g. monologue 1 is meeting participant one speaking). The second set (lexicon 2, defined in [131]) comprehends the full first set, but also has combinations of two parallel actions (like a presentation and note-taking). The second set includes fourteen group actions. Both sets and a brief description are shown in table 17.

Table 17: Group action lexicon 1 and 2

Action	Lexicon	Description
Discussion	lexicon 1 and 2	most participants engaged in conversations
Monologue	lexicon 1 and 2	one participant speaking continuously without interruption
Monologue+ Note-taking	contained only in lexicon 2	one participant speaking continuously others taking notes
Note-taking	lexicon 1 and 2	most participants taking notes
Presentation	lexicon 1 and 2	one participant presenting using the projector screen
Presentation+ Note-taking	contained only in lexicon 2	one participant presenting using projector screen, others taking notes
White-board	lexicon 1 and 2	one participant speaking using the white-board
White-board+ Note-taking	contained only in lexicon 2	one participant speaking using white-board, others taking notes

11.2 Meeting Data Set

We used a public corpus of 59 five-minute, four-participant scripted meetings ([75]). The recordings took place at IDIAP in an instrumented meeting room equipped with cameras and microphones[20]. Video has been recorded using 3 fixed cameras. Two cameras capture a frontal view of the meeting participants, and the third camera captures the white-board and the projector screen. Audio was recorded using lapel microphones attached to participants, and an eight-microphone array placed in the centre of the table.

[20] This corpus is publicly available from http://mmm.idiap.ch/

11.3 Features

The investigated individual actions are multimodal, we therefore use different audio-visual features. They are distinguished between *person-specific* AV features and *group-level* AV features. The former are extracted from individual participants. The latter are extracted from the white-board and projector screen regions. Furthermore we use a small set of lexical features. The features are described in the next paragraphs, for details please refer to the indicated literature.

From the large number of available features each site has chosen a set, used to train and evaluate their models. The complete list of features, and the three different sets IDIAP, TUM, UEDIN are listed in table 18.

11.3.1 Audio features

MFCC: For each of the speakers four MFC coefficients and the energy were extracted from the lapel-microphones. This results in a 20-dimensional vector $\vec{x}_S(t)$ containing speaker-dependent information.

A binary speech and silence segmentation (BSP) for each of the six locations in the meeting room was extracted with the SRP-PHAT measure ([75]) from the microphone array. This results in a six-dimensional discrete vector $\vec{x}_{BSP}(t)$ containing position dependent information.

Prosodic features are based on a denoised and stylised version of the intonation contour, an estimate of the syllabic rate of speech and the energy ([37]). These acoustic features comprise a 12 dimensional feature vector (3 features for each of the 4 speakers).

Speaker activity features rely on the active speaker locations evaluated using a sound source localisation process based on a microphone array ([75]). A 216 element feature vector resulted from all the 6^3 possible products of the 6 most probable speaker locations (four seats and two presentation positions) during the most recent three frames ([37]). A speaker activity feature vector at time t thus gives a local sample of the speaker interaction pattern in the meeting at around time t.

Further audio features: From the microphone array signals, we first compute a speech activity measure (SRP-PHAT). Three acoustic features, namely energy, pitch and speaking rate, were estimated on speech segments, zeroing silence segments. We used the SIFT algorithm to extract pitch, and a combination of estimators to extract speaking rate ([75]).

11.3.2 Global motion visual features

In the meeting room the four persons are expected to be at one of six different locations: one of four chairs, the whiteboard, or at a presentation position. For each location L in the meeting room a difference image sequence $I_d^L(x,y)$ is calculated by subtracting the pixel values of two subsequent frames from the video stream. Then seven global motion features ([133]) are derived from the image sequence: the centre of motion is calculated for the x- and y-direction, the changes in motion are used to express the dynamics of movements. Furthermore the mean absolute deviation of the pixels relative to the centre of motion is computed. Finally the intensity of motion is calculated from the average absolute value of the motion distribution. These seven features are concatenated for each time step in the location dependent motion vector. Concatenating the motion vectors from each of the six positions leads to the final visual feature vector that describes the overall motion in the meeting room with 42 features.

Table 18: Audio, visual and semantic features, and the resulting three feature sets.

		Description	IDIAP	TUM	UEDIN
Person-Specific Features	Visual	head vertical centroid	X		
		head eccentricity	X		
		right hand horizontal centroid	X		
		right hand angle	X		
		right hand eccentricity	X		
		head and hand motion	X		
		global motion features from each seat		X	
	Audio	SRP-PHAT from each seat	X		
		speech relative pitch	X		X
		speech energy	X	X	X
		speech rate	X		X
		4 MFCC coefficients		X	
		binary speech and silence segmentation		X	
	Semantic	individual gestures		X	
		talking activity		X	
Group Features	Visual	mean difference from white-board	X		
		mean difference from projector screen	X		
		global motion features from whiteboard		X	
		global motion features from projector screen		X	
	Audio	SRP-PHAT from white-board	X		
		SRP-PHAT from projector screen	X		
		speaker activity features			X
		binary speech from white-board		X	
		binary speech from projector screen		X	

11.3.3 Skin-colour blob visual features

Visual features derived from head and hands skin-colour blobs were extracted from the three cameras. For the two cameras looking at people, visual features extracted consist of head vertical centroid position and eccentricity, hand horizontal centroid position, eccentricity, and angle. The motion magnitude for head and hand blobs were also extracted. The average intensity of difference images computed by background subtraction was extracted from the third camera. All features were extracted at 5 frames per second, and the complete set of features is listed in table 18. For details refer to [131].

11.3.4 Semantic features

Originating from the low level features also features on a higher level have been extracted. For each of the six locations in the meeting room the talking activity has been detected using results from [69]. Further individual gestures of each participant have been detected using the gesture recogniser from [133]. The possible features were all normalised to the length of the meeting event to provide the relative duration of this particular feature. From all available events only those that are highly discriminative were chosen which resulted in a nine dimensional feature vector.

11.4 Models for Group Action Segmentation and Recognition

11.4.1 Meeting segmentation using semantic features

This approach combines the detection of the boundaries and classification of the segments in one step. The strategy is similar to that one used in the BIC-Algorithm ([113]). Two connected windows with variable length are shifted

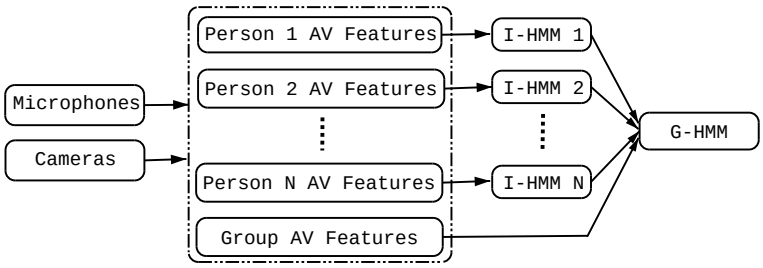

Figure 11: Multi-layer HMM on group action recognition.

over the time scale. Thereby the inner border is shifted from the left to the right in steps of one second and in each window the feature vector is classified by a low-level classifier. If there is a different result in the two windows, the inner border is considered a boundary of a meeting event. If no boundary is detected in the actual window, the whole window is enlarged and the inner border is again shifted from left to the right. Details can be found in [95].

11.4.2 Multi-stream mixed-state DBN for disturbed data

In real meetings the data can be disturbed in various ways: events like slamming of a door may mask the audio channel or background babble may appear; the visual channel can be (partly) masked by persons standing or walking in front of a camera. We therefore developed a novel approach for meeting event recognition, based on mixed-state DBNs, that can handle noise and occlusions in all channels ([4, 5]). Mixed-state DBNs are an HMM coupled with a LDS, they have been applied to recognising human gestures in [86]. Here, this approach has been extended to a novel multi-stream DBN for meeting event recognition.

Each of the three observed features: microphone array (BSP), lapel microphone (MFCC) and the visual global motion stream (GM) is modelled in a separate stream. The streams correspond to a multi-stream HMM, where each stream has a separate representation for the features. However, the visual stream is connected to a LDS, resulting in a mixed-state DBN. Here the LDS is a Kalman filter, using information from all streams as driving input, to smooth the visual stream. With this filtering, movements are predicted based on the previous time-slice and on the state of the multi-stream HMM at the current time. Thus occlusions can be compensated with the information from all channels. Given an observation O and the model parameters E_j for the mixed-state DBN, the joint probability of the model is the product of the stream probabilities: $P(O,E_j) = P_B \cdot P_M \cdot P_G$. The model parameters are learned for each of the eight event classes j with a variational learning EM-algorithm during the training phase. During the classification an unknown observation O is presented to all models E_j. Then $P(O|E_j)$ is calculated for each model and O is assigned to the class with the highest likelihood: $\text{argmax}_{E_j \in E} P(O|E_j)$. Applying the Viterbi-algorithm to the model, leads to a meeting event segmentation framework. The mixed-state DBN can therefore easily be combined with other models presented in this document.

11.4.3 Multi-layer Hidden Markov Model

In this section we summarise the multi-layer HMM applied to group action recognition. For a detailed discussion, please refer to [131].

In the multi-layer HMM framework, we distinguish group actions (which belong to the whole set of participants, such as *discussion and presentation*) from individual actions (belonging to specific persons, such as *writing and speaking*). To recognise group actions, individual actions act as the bridge between group actions and low-level features, thus decomposing the problem in stages, and simplifying the complexity of the task.

Let I-HMM denote the lower recognition layer (individual action), and G-HMM denote the upper layer (group action). I-HMM receives as input audio-visual (AV) features extracted from each participant, and outputs posterior

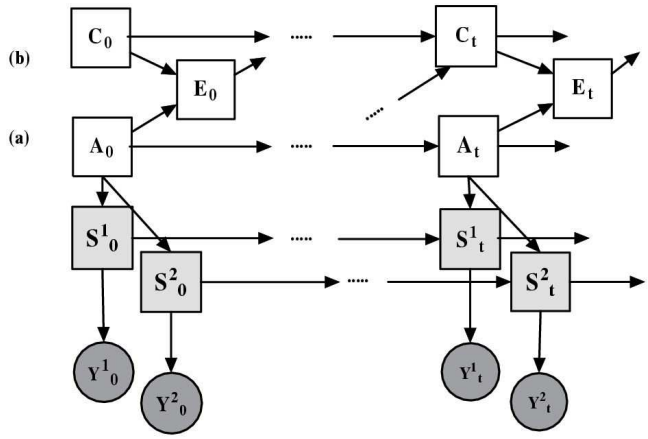

Figure 12: Multistream DBN model (a) enhanced with a "counter structure" (b); square nodes represent discrete hidden variables and circles must be intend as continuous observations

probabilities of the individual actions given the current observations. In turn, G-HMM receives as input the output from I-HMM, and a set of group features, directly extracted from the raw streams, which are not associated to any particular individual. In the multi-layer HMM framework, each layer is trained independently, and can be substituted by any of the HMM variants that might capture better the characteristics of the data, more specifically asynchrony ([16]), or different noise conditions between the audio and visual streams ([38]). The multi-layer HMM framework is summarised in figure 11.

Compared with a single-layer HMM, the layered approach has the following advantages, some of which were previously pointed out by [82]: (1) a single-layer HMM is defined on a possibly large observation space, which might face the problem of over-fitting with limited training data. It is important to notice that the amount of training data becomes an issue in meetings where data labelling is not a cheap task. In contrast, the layers in our approach are defined over small-dimensional observation spaces, resulting in more stable performance in cases of limited amount of training data. (2) The I-HMMs are person-independent, and in practice can be trained with much more data from different persons, as each meeting provides multiple individual streams of training data. Better generalisation performance can then be expected. (3) The G-HMMs are less sensitive to slight changes in the low-level features because their observations are the outputs of the individual action recognisers, which are expected to be well trained. (4) The two layers are trained independently. Thus, we can explore different HMM combination systems. In particular, we can replace the baseline I-HMMs with models that are more suitable for multi-modal asynchronous data sequences. The framework thus becomes simpler to understand, and amenable to improvements at each separate level.

11.4.4 Multistream DBN model

The DBN formalism allows the construction and development of a variety of models, starting from a simple HMM and extending to more sophisticated models (hierarchical HMMs, coupled HMMs, etc). With a small effort, DBNs are able to factorise the internal state space, organising it in a set of interconnected and specialised hidden variables.

Our multi-stream model (bottom of figure 12) exploits this principle in two ways: decomposing meeting actions into smaller logical units, and modelling parallel feature streams independently. We assume that a meeting action can be decomposed into a sequence of small units: meeting subactions. In accordance with this assumption the state space is decomposed into two levels of resolution: meeting actions (nodes A) and meeting subactions (nodes S^F). Note that the decomposition of meeting actions into meeting subactions is done automatically through the training process.

Feature sets derived from different modalities are usually governed by different laws, have different character-

istic time-scales and highlight different aspects of the communicative process. Starting from this hypothesis we further subdivided the model state space according to the nature of features that are processed, modelling each feature stream independently (multistream approach). The resulting model has an independent substate node S^F for each feature class F , and integrates the information carried by each feature stream at a 'higher level' of the model structure (arcs between A and S^F,$F = [1,n]$). Since the adopted *lexicon 1* (section 11.1) is composed by 8 meeting actions, the action node A has a cardinality of 8. The cardinalities of the sub-action nodes S are part of parameter set, and in our experiments we have chosen a value of 6 or 7.

The probability to remain in an HMM state corresponds to an inverse exponential ([88]): a similar behaviour is displayed by the proposed model. This distribution is not well-matched to the behaviour of meeting action durations. Rather than adopting ad hoc solutions, such as action transition penalties, we preferred to improve the flexibility of state duration modelling, by enhancing the existing model with a counter structure (top of figure 12). The counter variable C, which is ideally incremented during each action transition, attempts to model the expected number of recognised actions. Action variables A now also generate the hidden sequence of counter nodes C, together with the sequence of sub-action nodes S. Binary enabler variables E have an interface role between action variables A and counter nodes C.

This model presents several advantages over a simpler HMM in which features are "early integrated" into a single feature vector: feature classes are processed independently according to their nature; more freedom is allowed in the state space partitioning and in the optimisation of the sub-state space assigned to each feature class; knowledge from different streams is integrated together at an higher level of the model structure; etc. Unfortunately all these advantages, and the improved accuracy that can be achieved, are balanced by an increased model size, and therefore by an increased computational complexity.

11.5 Performance Measures

Since group meeting actions are high level symbols and their boundaries are extremely vague. In order to evaluate results of the segmentation and recognition task we used the Action Error Rate, a metric that privileges the recognition of the correct action sequence, rather than the precise temporal boundaries. AER is defined as the sum of *insertion* (Ins), *deletion* (Del), and *substitution* (Subs) errors, divided by the total number of actions in the ground-truth:

$$\mathrm{AER} = \frac{\mathrm{Subs} + \mathrm{Del} + \mathrm{Ins}}{\mathrm{Total\ Actions}} \times 100\% \tag{8}$$

Measures based on *deletion* (Del) and *insertion* (Ins) and *substitution* (Subs) are also used to evaluate action recognition results.

11.6 Experiments and Discussions

11.6.1 Higher semantic feature approach

The results of the segmentation are shown in table 19 (BN: Bayesian Network, GMM: Gaussian Mixture Models, MLP: Multilayer Perceptron Network, RBF: Radial Basis Network, SVM: Support Vector Machines). Each row denotes the classifier that was used. The columns show the insertion rate (number of insertions in respect to all meeting events), the deletion rate (number of deletions in respect to all meeting events), the accuracy (mean absolute error) of the found segment boundaries in seconds and the recognition error rate. In all columns lower numbers denote better results. As can be seen from the tables, the results are quite variable and heavily depend on the used classifier. These results are comparable to the ones presented in [94]. With the integrated approach the best outcome is achieved by the radial basis network. Here the insertion rate is the lowest. The detected segment boundaries match pretty well with a deviation of only about five seconds to the original defined boundaries.

Table 19: Segmentation results using the higher semantic feature approach (BN: Bayesian Network, GMM: Gaussian Mixture Models, MLP: Multilayer Perceptron Network, RBF: Radial Basis Network, SVM: Support Vector Machines). The columns denote the insertion rate, the deletion rate, the accuracy in seconds and the classification error rate (using lexicon 1 in Table 17).

Classifier	Insertion (%)	Deletion (%)	Accuracy	Error (%)
BN	14.7	6.22	7.93	39.0
GMM	24.7	2.33	10.8	41.4
MLP	8.61	1.67	6.33	32.4
RBF	6.89	3.00	5.66	31.6
SVM	17.7	0.83	9.08	35.7

11.6.2 Multi-stream mixed-state DBN for disturbed data

To investigate the influence of disturbance to the recognition performance, the evaluation data was cluttered: the video data was occluded with a black bar covering one third of the image at different positions. The audio data from the lapel microphones and the microphone array was disturbed with a background-babble with 10dB SNR. 30 undisturbed videos were used for the training of the models. The remaining 30 unknown videos have been cluttered for the evaluation.

The novel DBN was compared to single-modal (audio and visual) HMMs, an early fusion HMM, and a multi-stream HMM. The DBN showed a significant improvement of the recognition rate for disturbed data. Compared to the baseline HMMs, the DBN reduced the recognition error by more than 1.5% (9% relative error reduction) for disturbed data. It may therefore be useful to combine this approach with the other models presented in this document, to improve the noise robustness. Please refer to [4, 5] for detailed recognition results, as well as a comprehensive description of the model.

11.6.3 Multi-layer hidden Markov model

Table 20 reports the performance in terms of action error rate (AER) for both multi-layer HMM and the single-layer HMM methods. Several configurations were compared, including audio-only, visual-only, early integration, multi-stream ([38]) and asynchronous HMMs ([16]). We can see that (1) the multi-layer HMM approach always outperforms the single-layer one, (2) the use of AV features always outperforms the use of single modalities for both single-layer and multi-layer HMM, supporting the hypothesis that the group actions we defined are inherently multimodel, (3) the best I-HMM model is the asynchronous HMM, which suggests that some asynchrony exists for our task of group action recognition, and is actually well captured by the asynchronous HMM.

11.6.4 Multistream DBN model

All the experiments depicted in this section were conducted on 53 meetings (subset of the meeting corpus depicted in section 11.2) using the lexicon 1 of eight group actions. We implemented the proposed DBN models using the Graphical Models Toolkit (GMTK) ([17]), and the evaluation is performed using a leave-one-out cross-validation procedure.

Table 21 shows experimental results achieved using: an ergodic 11-states HMM, a multi-stream approach (section 11.4.4) with two feature streams, and the full counter enhanced multi-stream model. The base 2-stream approach has been tested in two different sub-action configurations: imposing $\left|S^1\right| = \left|S^2\right| = \{6\,or\,7\}$. Therefore four experimental setups were investigated; and each setup has been tested with 3 different feature sets, leading to 12 independent experiments. The first feature configuration ("UEDIN") associates prosodic features and speaker activity features (section 11.3.1) respectively to the stream S^1 and to S^2. The feature configuration labelled as

Table 20: AER (%) for single-layer and multi-layer HMM (using lexicon 2 in Table 17).

Method		AER (%)
Single-layer HMM	Visual only	48.2
	Audio only	36.7
	Early Integration	23.7
	Mutli-stream	23.1
	Asynchronous	22.2
Multi-layer HMM	Visual only	42.4
	Audio only	32.3
	Early Integration	16.5
	Multi-stream	15.8
	Asynchronous	15.1

"IDIAP" makes use of the multimodal features extracted at IDIAP, representing audio related features (prosodic data and speaker localisation) through the observable node Y^1 and video related measures through Y^2. The last setup ("TUM") relies on two feature families extracted at the Technische Universität München: binary speech profiles derived from IDIAP speaker locations and video related global motion features; each of those has been assigned to an independent sub-action node. Note that in the HMM based experiment the only observable feature stream Y has been obtained by merging together both the feature vectors Y^1 and Y^2. Considering only the results (of table 21) obtained within the UEDIN feature setup, it is clear that the simple HMM shows much higher error than any other multi-stream configuration. The adoption of a multistream based approach reduces the AER to less than 20%, providing the lowest AER (11%) when sub-action cardinalities are fixed to 7. UEDIN features seem to provide a higher accuracy if compared with IDIAP and TUM setups, but it is essential to remember that our DBN models have been optimised for the UEDIN features. In particular sub-action cardinalities have been intensively studied with our features, but it will be interesting to discover optimal values for IDIAP and TUM features too. Moreover overall performances achieved with the multistream approach are very similar (AER are always in the range from 26.7% to 11.0%), and all my be considered promising. The TUM setup seems to be the configuration for which switching from a HMM to a multistream DBN approach provides the greatest improvement in performance: the error rate decreases from 92.9% to 21.4%. If with the UEDIN feature set the adoption of a counter structure is not particularly effective, with IDIAP features the counter provides a significant AER reduction (from 26.7% to 24.9%). We are confident that further improvements with IDIAP features could be obtained by using more than 2 streams (such as the 3 multistream model adopted in [37]). Independently of the feature configuration, the best overall results are achieved with the multistream approach and a state space of 7 by 7 substates.

11.7 Summary and conclusions

We have presented the joint efforts of three institutes (TUM, IDIAP and UEDIN) towards structuring meetings into sequences of multimodal human interactions. A large number of different audio-visual features have been extracted from a common meeting data corpus. From this features, three multimodal sets have been chosen. Four different frameworks towards automatic segmentation and classification of meetings into action units haven been proposed.

The first approach from TUM exploits higher semantic features for structuring a meeting into group actions. It thereby uses an algorithm that is based on the idea of the Bayesian-Information-Criterion. The mixed-state DBN approach developed by TUM compensates for disturbances in both the visual and the audio channel. It is not a segmentation framework but can be integrated into the other approaches presented in this section to improve their robustness. The multi-layer Hidden Markov Model developed by IDIAP decomposes group actions as a two-layer process, one that models basic individual activities from low-level audio-visual features, and another one that models the group action (belonging to the whole set of participants). The multi-stream DBN model proposed by

Table 21: AER (%) for an HMM, and for a multi-stream (2 streams) approach with and without the "counter structure"; the models have been individually tested with the 3 different feature sets (using lexicon 1 in Table 17)

Model	Feature Set	Corr.	Sub.	Del.	Ins.	AER
HMM	UEDIN	63.3	13.2	23.5	11.7	48.4
	IDIAP	62.6	19.9	17.4	24.2	61.6
	TUM	60.9	25.6	13.5	53.7	92.9
2 streams $\left(\left\|S^F\right\| = 6\right)$	UEDIN	86.1	5.7	8.2	3.2	17.1
	IDIAP	77.9	8.9	13.2	4.6	26.7
	TUM	85.4	9.3	5.3	6.8	21.4
2 streams $\left(\left\|S^F\right\| = 6\right)$ + counter	UEDIN	85.8	7.5	6.8	4.6	18.9
	IDIAP	79.4	10.0	10.7	4.3	24.9
	TUM	85.1	5.7	9.3	6.4	21.4
2 streams $\left(\left\|S^F\right\| = 7\right)$	UEDIN	90.7	2.8	6.4	1.8	11.0
	IDIAP	86.5	7.8	5.7	3.2	16.7
	TUM	82.9	7.1	10.0	4.3	21.4

UEDIN operates an unsupervised subdivision of meeting actions into sequences of group sub-actions, processing multiple asynchronous feature streams independently, introducing also a model extension to improve state duration modelling.

All presented approaches have provided comparable good performances, and there is still space for further improvements both in the feature domain (i.e.: exploit more modalities) and in the model infrastructure. Therefore in the near future we are going to investigate combinations of the proposed systems to improve the AER and to exploit the complementary strengths of the different approaches. Moreover the proposed approaches are easily generalizable to more elaborate segmentation and structuring tasks. Therefore it is our intention to adopt a richer set of "group meeting actions", and to validate the proposed frameworks on a more realist multimodal meeting corpus like the "AMI meeting corpus" ([27]), that is characterised by real, fully unconstrained meetings.

Another promising direction of research is action clustering, where typical activities can be identified on an unsupervised basis. Initial work in this direction was presented in [130]. Another direction for action recognition involves the use of partially labeled data. An initial approach was presented in [129].

12 Component Evaluation

For many of the areas covered in workpackage 5 and thus this document, we have devised component evaluation schemes and will perform the individual evaluations in the first half of 2006. The evaluation schemes will be published this fall as a first draft of deliverable 5.2. The full deliverable, due in month 30 of the project, will also contain the results of the component evaluations.

Currently, evaluation schemes for the following components are being defined:

- Topic Segmentation
- Meeting acts
- Dialog Acts & Segmentation ICSI
- Addressing
- Named Entities
- Extractive Summaries (on ICSI)
- Abstractive Summaries
- Indexing/Retrieval
- Chunking

13 Search engine for LVCSR-based keyword spotting in meeting data

13.1 Introduction

One of tasks of Brno University of Technology in AMI is to provide the project with keyword spotting (KWS) in meeting environment. We are working on several approaches to KWS including searching large vocabulary continuous speech recognition (LVCSR) lattices, acoustic search and a hybrid ”phonetic” search [3].

The most straightforward way to search in an output of LVCSR speech recognizer is to use existing search engines on the textual (“1-best”) output. We can however advantageously use a richer output of the recognizer – most recognition engines are able to produce an oriented graph of hypotheses called *lattice*. On contrary to 1-best output, the lattices tend to be complex and large. For efficient searching in such a complex and large data structure, the creation of an optimized indexing system which is the core of each fast search engine is necessary. The proposed system is based on principles used in Google [4]. It consists of indexer, sorter and searcher [5].

13.2 Input to the system

Word lattices generated by LVCSR are input to the indexing and search engine. The lattices (see example in Fig. 13) are stored in HTK standard lattice format (SLF) [7].

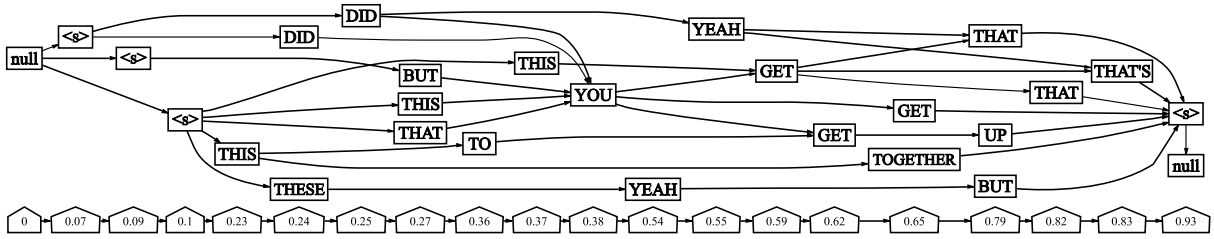

Figure 13: Example of a word lattice

13.3 The indexer

processes lattices stored in SLF files and stores them into system’s data structures. The indexing mechanism consists of three main phases:

- creating the lexicon
- storing and indexing lattices, creating the forward index
- creating the reverse index (based on the forward index)

The lexicon provides a transformation from word to a unique number (ID) and vice versa. It saves the used disk space and also the time of comparing strings (number of bytes for storing numbers is less than the average length of word).

Lattices are stored in a structure which differs from the SLF structure. For each search result it is needed not only to show the time of found word, but also its context. It means that we need to traverse the lattice from the found word in both directions (forward and backward) to gather those words lying on the best path which traverses through the found word. On contrary to SLF, where nodes are separated from links, lattices are converted to another structure which stores all forward and backward links for each particular node at one place. It is also needed to assign *a confidence* to each hypothesis. This is given by the log-likelihood ratio:

$$C^{lvcsr}(KW) = L^{lvcsr}_{alpha}(KW) + L^{lvcsr}(KW) + L^{lvcsr}_{beta}(KW) - L^{lvcsr}_{best}, \quad (9)$$

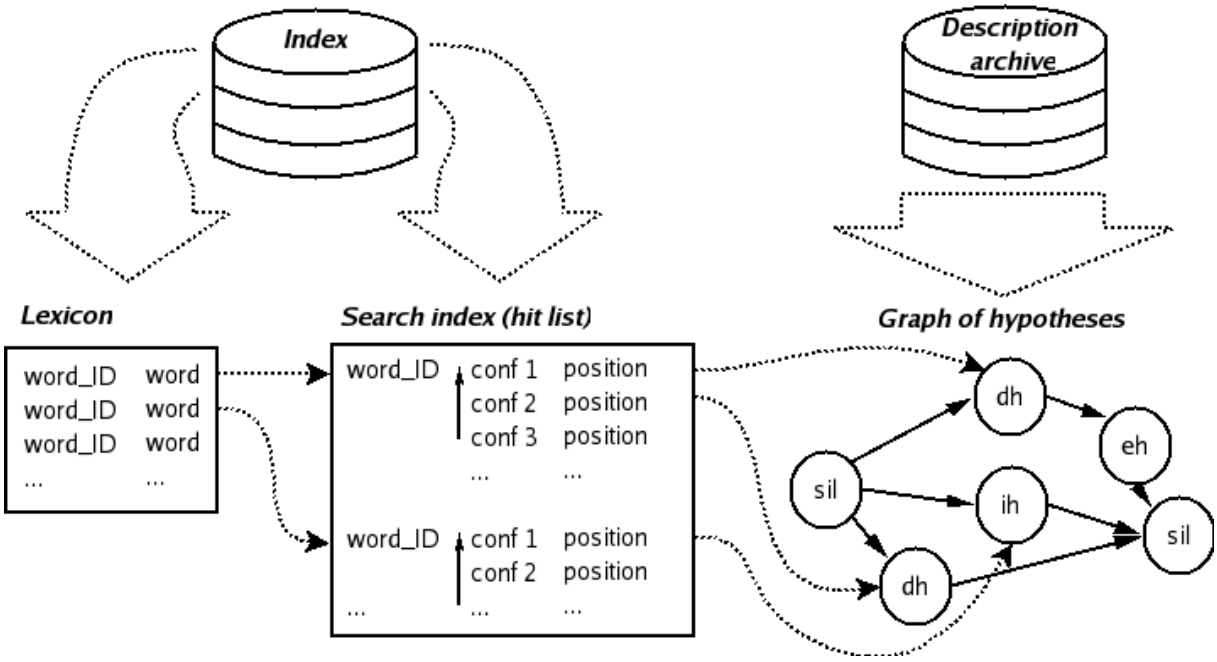

Figure 14: Simplified index structure

where the forward likelihood $L_{alpha}^{lvcsr}(KW)$ is the likelihood of the best path through lattice from the beginning of lattice to the keyword and backward likelihood $L_{beta}^{lvcsr}(KW)$ is computed from the end of lattice to the keyword. These two likelihoods are computed by the standard Viterbi formulae:

$$L_{alpha}^{lvcsr}(N) = L_{a}^{lvcsr}(N) + L_{l}^{lvcsr}(N) + \min_{N_P} L_{alpha}^{lvcsr}(N_P) \tag{10}$$

$$L_{beta}^{lvcsr}(N) = L_{a}^{lvcsr}(N) + L_{l}^{lvcsr}(N) + \min_{N_F} L_{beta}^{lvcsr}(N_F) \tag{11}$$

where N_F is set of nodes directly following node N (nodes N and N_F are connected by an arc), N_P is set of nodes directly preceding node N. $L_{a}^{lvcsr}(N)$ and $L_{l}^{lvcsr}(N)$ are acoustic and language-model likelihoods respectively.

The algorithm is initialized by setting $L_{alpha}^{lvcsr}(first) = 0$ and $L_{beta}^{lvcsr}(last) = 0$. The last likelihood we need in Eq. 9: $L_{best}^{lvcsr} = L_{alpha}^{lvcsr} = L_{beta}^{lvcsr}$ is the likelihood of the most probable path through the lattice.

While processing lattices, the indexer stores each hypothesis into the forward index, so that the forward index is sorted by `documentID` and by time. Such index can be useful for searching in some particular document, but for global searching we need a reverse index [4].

13.4 The sorter

During the phase of indexing and storing lattices, the forward index is created. It stores each hypothesis (word, it's confidence, time and position in lattice file) from lattice into a hit list. Records in the forward index are sorted by `documentID` (number which represents the lattice's file name) and time. The forward index itself is however not very useful for searching for a particular word, because it would be necessary to go through the hit list sequentially and select only matching words. Therefore the reverse index is created (like in Google) which has the same structure as the forward index, but is sorted by words and by confidence of hypotheses. It means that all

occurrences of a particular word are stored at one place. There is also a table which transforms any word from lexicon into the start position of corresponding list in reverse index.

Searching for one word then consists only in jumping right to the beginning of it's list in reverse index, selecting first few occurrences and getting their context from corresponding lattice. The advantage of splitting the indexing mechanism into three phases is that the second phase (storing and indexing lattices), which is the most CPU time consuming one, can be run in parallel on several computers. Each parallel process creates its own forward index. These indices are then merged together and sorted to create the reverse index.

13.5 The searcher

uses the reverse index to find occurrences of words from query and then it discovers whether they match the whole query or not. For all matching occurrences, it loads into the memory only a small part of lattice within which the found word occurs. Then the searcher traverses this part of lattice in forward and backward direction selecting only the best hypotheses; in this way it creates the most probable string traversing the found word.

13.6 Experiment

The system was tested on four AMI pilot meetings, each with four speakers and total duration of about 1.9 hours. The recognition lattices were generated using the AMI-LVCSR system incorporating state-of-the-art acoustic and language modeling techniques [6].

For testing data of 1.9 hour, the lattices consist of 3,607,089 hypotheses and 36,036,967 arcs. Searching and looking for the context of 6 hypotheses takes about 3 seconds. Although the system is not yet well-optimized, it produces search results quite fast. Approximately 95% of time is spent on looking for the context of the found word. It is possible to optimize this process with expected increase of speed by 70-80%.

13.7 Conclusions

We have presented a system for fast search in speech recognition lattices making extensive use of indexing. The results obtained with this system are promising, the software has been integrated with the meeting browser JFerret [2] and presented at several occasions. Currently, we are testing the extension of the system allowing to enter multi-word queries, and options to narrow search space (limitation only to particular meetings, speakers, time intervals). We also plan to employ this system in phoneme-lattice based keyword spotting which eliminates the main drawback of LVCSR — the dependency on recognition vocabulary [3].

References

[1] Marc Al-Hames et al.: *D4.1 Report on Implementation of Audio, Video, and Multimodal Algorithms*, AMI deliverable, December 2004.

[2] Bram van der Wal et al.: *D6.3 Preliminary demonstrator of Browser Components and Wireless Presentation System*, AMI deliverable, August 2005, in preparation.

[3] Igor Szöke, Petr Schwarz, Pavel Matějka, Lukáš Burget, Martin Karafiát, Michal Fapšo, Jan Černocký: *Comparison of Keyword Spotting Approaches for Informal Continuous Speech*, accepted to Eurospeech 2005.

[4] Sergey Brin, Lawrence Page: *The Anatomy of a Large-Scale Hypertextual Web Search Engine, Computer Science Department*, Stanford University (2002)

[5] Michal Fapso, Petr Schwarz, Igor Szoke, Lukas Burget, Martin Karafiat, Jan Cernocky: *Search Engine for Information Retrieval from Multi-modal Records*, poster at 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, Edinburgh, July 2005.

[6] Thomas Hain, Lukas Burget, John Dines, Giulia Garau, Martin Karafiat, Mike Lincoln, Iain McCowan, Darren Moore, Vincent Wan, Rolland Ordelman, and Steve Renals: *The 2005 AMI System for the Transcription of Speech in Meetings*, in Proc. NIST Rich Transcription 2005 Spring Meeting Recognition Evaluation, Edinburgh, July 2005

[7] Steve Young et al.: *The HTK Book (for HTK Version 3.3)*, Cambridge University Engineering Department, April 2005.

14 Extractive Summaries

Additional work on extractive summarization by Buist, Kraaij and Raaijmakers ist published in [25].

14.1 Introduction

In the field of automatic summarization, it is widely agreed upon that more attention needs to be paid to the development of standardized approaches to summarization evaluation. For example, the current incarnation of the Document Understanding Conference is putting its main focus on the development of evaluation schemes, including semi-automatic approaches to evaluation. One semi-automatic approach to evaluation is ROUGE [70], which is primarily based on n-gram co-occurrence between automatic and human summaries. A key question of the research contained herein is how well ROUGE correlates with human judgments of summaries within the domain of meeting speech. If it is determined that the two types of evaluations correlate strongly, then ROUGE will likely be a valuable and robust evaluation tool in the development stage of a summarization system, when the cost of frequent human evaluations would be prohibitive.

Three basic approaches to summarization are evaluated and compared below: Maximal Marginal Relevance, Latent Semantic Analysis, and feature-based classification. The other major comparisons in this paper are between summaries on ASR versus manual transcripts, and between manual and automatic extracts. For example, regarding the former, it might be expected that summaries on ASR transcripts would be rated lower than summaries on manual transcripts, due to speech recognition errors. Regarding the comparison of manual and automatic extracts, the manual extracts can be thought of as a gold standard for the extraction task, representing the performance ceiling that the automatic approaches are aiming for.

More detailed descriptions of the summarization approaches and experimental setup can be found in [78]. That work relied solely on ROUGE as an evaluation metric, and this paper proceeds to investigate whether ROUGE alone is a reliable metric for our summarization domain, by comparing the automatic scores with recently-gathered human evaluations.

14.2 Description of the Summarization Approaches

14.2.1 Maximal Marginal Relevance (MMR)

MMR [26] uses the vector-space model of text retrieval and is particularly applicable to query-based and multi-document summarization. The MMR algorithm chooses sentences via a weighted combination of query-relevance and redundancy scores, both derived using cosine similarity. The MMR score $Sc^{MMR}(i)$for a given sentence S_i in the document is given by

$$Sc^{MMR}(i) = \lambda(Sim(S_i, D)) - (1-\lambda)(Sim(S_i, Summ)),$$

where D is the average document vector, *Summ* is the average vector from the set of sentences already selected, and λ trades off between relevance and redundancy. *Sim* is the cosine similarity between two documents.

This implementation of MMR uses lambda annealing so that relevance is emphasized while the summary is still short and minimizing redundancy is prioritized more highly as the summary lengthens.

14.2.2 Latent Semantic Analysis (LSA)

LSA is a vector-space approach which involves projecting the original term-document matrix to a reduced dimension representation. It is based on the singular value decomposition (SVD) of an $m \times n$ term-document matrix A, whose elements A_{ij} represent the weighted term frequency of term i in document j. In SVD, the term-document matrix is decomposed as follows:

$$A = USV^T$$

where U is an $m \times n$ matrix of left-singular vectors, S is an $n \times n$ diagonal matrix of singular values, and V is the $n \times n$ matrix of right-singular vectors. The rows of V^T may be regarded as defining topics, with the columns

representing sentences from the document. Following Gong and Liu [48], summarization proceeds by choosing, for each row in V^T, the sentence with the highest value. This process continues until the desired summary length is reached.

Two drawbacks of this method are that dimensionality is tied to summary length and that good sentence candidates may not be chosen if they do not "win" in any dimension [105]. The authors in [105] found one solution, by extracting a single LSA-based sentence score, with variable dimensionality reduction.

We address the same concerns, following the Gong and Liu approach, but rather than extracting the best sentence for each topic, the *n* best sentences are extracted, with *n* determined by the corresponding singular values from matrix *S*. The number of sentences in the summary that will come from the first topic is determined by the percentage that the largest singular value represents out of the sum of all singular values, and so on for each topic. Thus, dimensionality reduction is no longer tied to summary length and more than one sentence per topic can be chosen. Using this method, the level of dimensionality reduction is essentially learned from the data.

14.2.3 Feature-Based Approaches

Feature-based classification approaches have been widely used in text and speech summarization, with positive results [67]. In this work we combined textual and prosodic features, using Gaussian mixture models for the extracted and non-extracted classes. The prosodic features were the mean and standard deviation of F0, energy, and duration, all estimated and normalized at the word-level, then averaged over the utterance. The two lexical features were both TFIDF-based: the average and the maximum TFIDF score for the utterance.

For our second feature-based approach, we derived single LSA-based sentence scores [105] to complement the six features described above, to determine whether such an LSA sentence score is beneficial in determining sentence importance. We reduced the original term-document matrix to 300 dimensions; however, Steinberger and Ježek found the greatest success in their work by reducing to a single dimension (Steinberger, personal communication). The LSA sentence score was obtained using:

$$Sc_i^{LSA} = \sqrt{\sum_{k=1}^{n} v(i,k)^2 * \sigma(k)^2},$$

where $v(i,k)$ is the *k*th element of the *i*th sentence vector and $\sigma(k)$ is the corresponding singular value.

14.3 Experimental Setup

We used human summaries of the ICSI Meeting corpus for evaluation and for training the feature-based approaches. An evaluation set of six meetings was defined and multiple human summaries were created for these meetings, with each test meeting having either three or four manual summaries. The remaining meetings were regarded as training data and a single human summary was created for these. Our summaries were created as follows.

Annotators were given access to a graphical user interface (GUI) for browsing an individual meeting that included earlier human annotations: an orthographic transcription time-synchronized with the audio, and a topic segmentation based on a shallow hierarchical decomposition with keyword-based text labels describing each topic segment. The annotators were told to construct a textual summary of the meeting aimed at someone who is interested in the research being carried out, such as a researcher who does similar work elsewhere, using four headings:

- general abstract: "why are they meeting and what do they talk about?";
- decisions made by the group;
- progress and achievements;
- problems described

The annotators were given a 200 word limit for each heading, and told that there must be text for the general abstract, but that the other headings may have null annotations for some meetings.

Immediately after authoring a textual summary, annotators were asked to create an extractive summary, using a different GUI. This GUI showed both their textual summary and the orthographic transcription, without topic segmentation but with one line per dialogue act based on the pre-existing MRDA coding [101] (The dialogue act categories themselves were not displayed, just the segmentation). Annotators were told to extract dialogue acts that together would convey the information in the textual summary, and could be used to support the correctness of that summary. They were given no specific instructions about the number or percentage of acts to extract or about redundant dialogue act. For each dialogue act extracted, they were then required in a second pass to choose the sentences from the textual summary supported by the dialogue act, creating a many-to-many mapping between the recording and the textual summary.

The MMR and LSA approaches are both unsupervised and do not require labelled training data. For both feature-based approaches, the GMM classifiers were trained on a subset of the training data representing approximately 20 hours of meetings.

We performed summarization using both the human transcripts and speech recognizer output. The speech recognizer output was created using baseline acoustic models created using a training set consisting of 300 hours of conversational telephone speech from the Switchboard and Callhome corpora. The resultant models (cross-word triphones trained on conversational side based cepstral mean normalised PLP features) were then MAP adapted to the meeting domain using the ICSI corpus [53]. A trigram language model was employed. Fair recognition output for the whole corpus was obtained by dividing the corpus into four parts, and employing a leave one out procedure (training the acoustic and language models on three parts of the corpus and testing on the fourth, rotating to obtain recognition results for the full corpus). This resulted in an average word error rate (WER) of 29.5%. Automatic segmentation into dialogue acts or sentence boundaries was not performed: the dialogue act boundaries for the manual transcripts were mapped on to the speech recognition output.

14.3.1 Description of the Evaluation Schemes

A particular interest in our research is how automatic measures of informativeness correlate with human judgments on the same criteria. During the development stage of a summarization system it is not feasible to employ many hours of manual evaluations, and so a critical issue is whether or not software packages such as ROUGE are able to measure informativeness in a way that correlates with subjective summarization evaluations.

ROUGE Gauging informativeness has been the focus of automatic summarization evaluation research. We used the ROUGE evaluation approach [70], which is based on n-gram co-occurrence between machine summaries and "ideal" human summaries. ROUGE is currently the standard objective evaluation measure for the Document Understanding Conference [21]; ROUGE does not assume that there is a single "gold standard" summary. Instead it operates by matching the target summary against a set of reference summaries. ROUGE-1 through ROUGE-4 are simple n-gram co-occurrence measures, which check whether each n-gram in the reference summary is contained in the machine summary. ROUGE-L and ROUGE-W are measures of common subsequences shared between two summaries, with ROUGE-W favoring contiguous common subsequences. Lin [70] has found that ROUGE-1 and ROUGE-2 correlate well with human judgments.

Human Evalautions The subjective evaluation portion of our research utilized 5 judges who had little or no familiarity with the content of the ICSI meetings. Each judge evaluated 10 summaries per meeting, for a total of sixty summaries. In order to familiarize themselves with a given meeting, they were provided with a human abstract of the meeting and the full transcript of the meeting with links to the audio. The human judges were instructed to read the abstract, and to consult the full transcript and audio as needed, with the entire familiarization stage not to exceed 20 minutes.

[21] http://duc.nist.gov/

The judges were presented with 12 questions at the end of each summary, and were instructed that upon beginning the questionnaire they should not reconsult the summary itself. 6 of the questions regarded informativeness and 6 involved readability and coherence, though our current research concentrates on the informativeness evaluations. The evaluations used a Likert scale based on agreement or disagreement with statements, such as the following Informativeness statements:

1. The important points of the meeting are represented in the summary.
2. The summary avoids redundancy.
3. The summary sentences on average seem relevant.
4. The relationship between the importance of each topic and the amount of summary space given to that topic seems appropriate.
5. The summary is repetitive.
6. The summary contains unnecessary information.

Statements such as 2 and 5 above are measuring the same impressions, with the polarity of the statements merely reversed, in order to better gauge the reliability of the answers. The readability/coherence portion consisted of the following statements:

1. It is generally easy to tell whom or what is being referred to in the summary.
2. The summary has good continuity, i.e. the sentences seem to join smoothly from one to another.
3. The individual sentences on average are clear and well-formed.
4. The summary seems disjointed.
5. The summary is incoherent.
6. On average, individual sentences are poorly constructed.

It was not possible in this paper to gauge how responses to these readability statements correlate with automatic metrics, for the reason that automatic metrics of readability and coherence have not been widely discussed in the field of summarization. Though subjective evaluations of summaries are often divided into informativeness and readability questions, only automatic metrics of informativeness have been investigated in-depth by the summarization community. We believe that the development of automatic metrics for coherence and readability should be a high priority for researchers in summarization evaluation and plan on pursuing this avenue of research. For example, work on coherence in NLG [68] could potentially inform summarization evaluation. Mani [73] is one of the few papers to have discussed measuring summary readability automatically.

14.4 Results

The results of these experiments can be analyzed in various ways: significant differences of ROUGE results across summarization approaches, deterioration of ROUGE results on ASR versus manual transcripts, significant differences of human evaluations across summarization approaches, deterioration of human evaluations on ASR versus manual transcripts, and finally, the correlation between ROUGE and human evaluations.

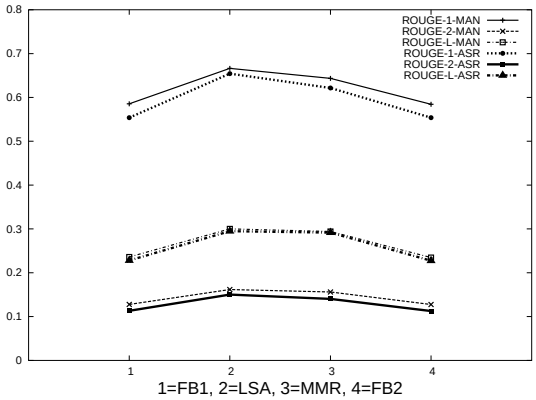

Figure 15: *ROUGE Scores for the Summarization Approaches*

14.4.1 ROUGE results across summarization approaches

All of the machine summaries were 10% of the original document length, in terms of the number of dialogue acts contained. Of the four approaches to summarization used herein, the latent semantic analysis method performed the best on every meeting tested for every ROUGE measure with the exception of ROUGE-3 and ROUGE-4. This approach was significantly better than either feature-based approach ($p<0.05$), but was not a significant improvement over MMR. For ROUGE-3 and ROUGE-4, none of the summarization approaches were significantly different from each other, owing to data sparsity. Figure 15 gives the ROUGE-1, ROUGE-2 and ROUGE-L results for each of the summarization approaches, on both manual and ASR transcripts.

ASR versus Manual The results of the four summarization approaches on ASR output were much the same, with LSA and MMR being comparable to each other, and each of them outperforming the feature-based approaches. On ASR output, LSA again consistently performed the best.

Interestingly, though the LSA approach scored higher when using manual transcripts than when using ASR transcripts, the difference was small and insignificant despite the nearly 30% WER of the ASR. All of the summarization approaches showed minimal deterioration when used on ASR output as compared to manual transcripts, but the LSA approach seemed particularly resilient, as evidenced by Figure 15. One reason for the relatively small impact of ASR output on summarization results is that for each of the 6 meetings, the WER of the summaries was lower than the WER of the meeting as a whole. Similarly, Valenza et al [114] and Zechner and Waibel [127] both observed that the WER of extracted summaries was significantly lower than the overall WER in the case of broadcast news. The table below demonstrates the discrepancy between summary WER and meeting WER for the six meetings used in this research.

Meeting	Summary WER/%	Meeting WER/%
Bed004	27.0	35.7
Bed009	28.3	39.8
Bed016	39.6	49.8
Bmr005	23.9	36.1
Bmr019	28.0	36.5
Bro018	25.9	35.6

WER Comparison for LSA Summaries and Meetings

There was no improvement in the second feature-based approach (adding an LSA sentence score) as compared with the first feature-based approach. The sentence score used here relied on a reduction to 300 dimensions, which may not have been ideal for this data.

STATEMENT	FB1	LSA	MMR	FB2
IMPORTANT POINTS	5.03	4.53	4.67	4.83
NO REDUNDANCY	**4.33**	2.60	3.00	3.77
RELEVANT	4.83	4.07	4.33	4.53
TOPIC SPACE	4.43	3.83	3.87	4.30
REPETITIVE	**3.37**	4.70	4.60	3.83
UNNECESSARY INFO.	**4.70**	6.00	5.83	5.00

Table 22: Human Ratings for 4 Approaches on Manual Transcripts

STATEMENT	FB1	LSA	MMR	FB2
IMPORTANT POINTS	3.53	**4.13**	3.73	3.50
NO REDUNDANCY	3.40	2.97	2.63	3.57
RELEVANT	3.47	3.57	3.00	3.47
TOPIC SPACE	3.27	3.33	3.00	3.20
REPETITIVE	4.43	4.73	4.70	4.20
UNNECESSARY INFO	5.37	6.00	6.00	5.33

Table 23: Human Ratings for 4 Approaches on ASR Transcripts

The similarity between the MMR and LSA approaches here mirrors Gong and Liu's findings, giving credence to the claim that LSA maximizes relevance and minimizes redundancy, in a different and more opaque manner then MMR, but with similar results. Regardless of whether or not the singular vectors of V^T can rightly be thought of as topics or concepts (a seemingly strong claim), the LSA approach was as successful as the more popular MMR algorithm.

14.4.2 Human results across summarization approaches

Table 14.4.2 presents average ratings for the six statements across four summarization approaches on manual transcripts. Interestingly, the first feature-based approach is given the highest marks on each criterion. For statements 2, 5 and 6 FB1 is significantly better than the other approaches. It is particularly surprising that FB1 would score well on statement 2, which concerns redundancy, given that MMR and LSA explicitly aim to reduce redundancy while the feature-based approaches are merely classifying utterances as relevant or not. The second feature-based approach was not significantly worse than the first on this score.

Considering the difficult task of evaluating ten extractive summaries per meeting, we are quite satisfied with the consistency of the human judges. For example, statements that were merely reworded versions of other statements were given consistent ratings. It was also the case that, with the exception of evaluating the sixth statement, judges were able to tell that the manual extracts were superior to the automatic approaches.

ASR versus Manual Table 14.4.2 presents average ratings for the six statements across four summarization approaches on ASR transcripts. The LSA and MMR approaches performed better in terms of having less deterioration of scores when used on ASR output instead of manual transcripts. LSA-ASR was not significantly worse than LSA on any of the 6 ratings. MMR-ASR was significantly worse than MMR on only 3 of the 6. In contrast, FB1-ASR was significantly worse than FB1 for 5 of the 6 approaches, reinforcing the point that MMR and LSA seem to favor extracting utterances with fewer errors. Figures 16, 17 and 18 depict how the ASR and manual approaches affect the INFORMATIVENESS-1, INFORMATIVENESS-4 and INFORMATIVENESS-6 ratings, respectively. Note that for Figure 6, a higher score is a worse rating.

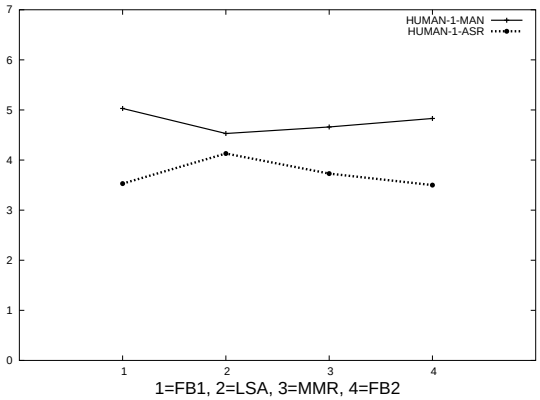

Figure 16: *INFORMATIVENESS-1 Scores for the Summarization Approaches*

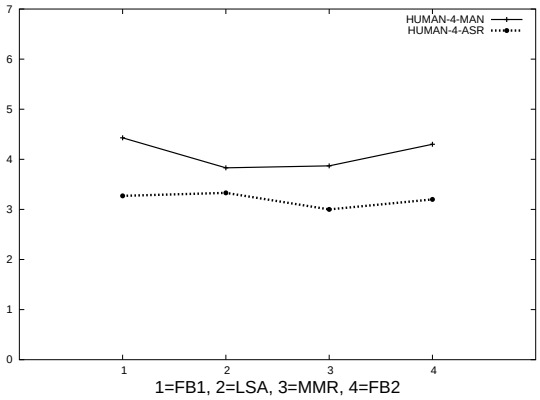

Figure 17: *INFORMATIVENESS-4 Scores for the Summarization Approaches*

14.4.3 ROUGE and Human correlations

According to [70], ROUGE-1 correlates particularly well with human judgments of informativeness. In the human evaluation survey discussed here, the first statement (INFORMATIVENESS-1) would be expected to correlate most highly with ROUGE-1, as it is asking whether the summary contains the important points of the meeting. As could be guessed from the discussion above, there is no significant correlation between ROUGE-1 and human evaluations when analyzing only the 4 summarization approaches on manual transcripts. However, when looking at the 4 approaches on ASR output, ROUGE-1 and INFORMATIVENESS-1 have a moderate and significant positive correlation (Spearman's rho = 0.500, $p < 0.05$). This correlation on ASR output is strong enough that when ROUGE-1 and INFORMATIVENESS-1 scores are tested for correlation across all 8 summarization approaches, there is a significant positive correlation (Spearman's rho = 0.388, $p < 0.05$).

The other significant correlations for ROUGE-1 across all 8 summarization approaches are with INFORMATIVENESS-2, INFORMATIVENESS-5 and INFORMATIVENESS-6. However, these are negative correlations. For example, with regard to INFORMATIVENESS-2, summaries that are rated as having a high level of redundancy are given high ROUGE-1 scores, and summaries with little redundancy are given low ROUGE-1 scores. Similary, with regard to INFORMATIVENESS-6, summaries that are said to have a great deal of unnecessary information are given high ROUGE-1 scores. It is difficult to interpret some of these negative correlations, as ROUGE does not measure redundancy and would not necessarily be expected to correlate with redundancy

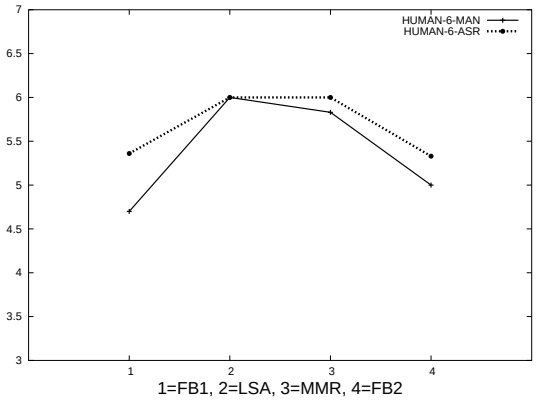

Figure 18: *INFORMATIVENESS-6 Scores for the Summarization Approaches*

evaluations.

14.5 Discussion

In general, ROUGE did not correlate well with the human evaluations for this data. The MMR and LSA approaches were deemed to be significantly better than the feature-based approaches according to ROUGE, while these findings were reversed according to the human evaluations. An area of agreement, however, is that the LSA-ASR and MMR-ASR approaches have a small and insignificant decline in scores compared with the decline of scores for the feature-based approaches. One of the most interesting findings of this research is that MMR and LSA approaches used on ASR tend to select utterances with fewer ASR errors.

ROUGE has been shown to correlate well with human evaluations in DUC, when used on news corpora, but the summarization task here – using conversational speech from meetings – is quite different from summarizing news articles. ROUGE may simply be less applicable to this domain.

14.6 Future Work

It remains to be determined through further experimentation by researchers using various corpora whether or not ROUGE truly correlates well with human judgments. The results presented above are mixed in nature, but do not present ROUGE as being sufficient in itself to robustly evaluate a summarization system under development.

We are also interested in developing automatic metrics of coherence and readability. We now have human evaluations of these criteria and are ready to begin testing for correlations between these subjective judgments and potential automatic metrics.

14.7 Acknowledgements

Thanks to Thomas Hain and the AMI-ASR group for the speech recognition output. This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811, publication).

15 Abstractive Summaries

The way a human summarizer abstracts a document is usually quite different from extractive summarization approaches. Where in the latter case, a machine applies statistical methods to select sentences regarded as relevant, a human summarizer usually proceeds in a less algorithmic way. Typically, she first reads[22] and *understands* the document, i.e. she builds up a mental model of the concepts and their interrelations in the source. In a second step (which is not necessarily performed sequentially, but may happen already by the time of reading), she abstracts this mental model to a condensed variant by leaving out parts considered irrelevant or by rearranging different parts of the model to yield simplified information structures. The resulting transformed mental model can be considered a *summarized* version of the original model, and a verbalization thereof would be the final summary.

15.1 ABSURD – Abstractive Summarization of Real-life Discourse

This rough and oversimplified description certainly falls short of the actual processes in a human mind. Yet, the above steps can provide a blueprint for an alternative to the extractive summarization approach. For AMI, DFKI *abstractive summarization group* designs the architecture of their summarization module ABSURD (see Figure 19), to resemble these different phases:

1. Understand
2. Abstract
3. Generate

In brevity, phase 1 is performed by the "discourse parser" component, phase 2 by the "information reduction & reorganization" component and phase 3 by the "document planner" and "realizer".

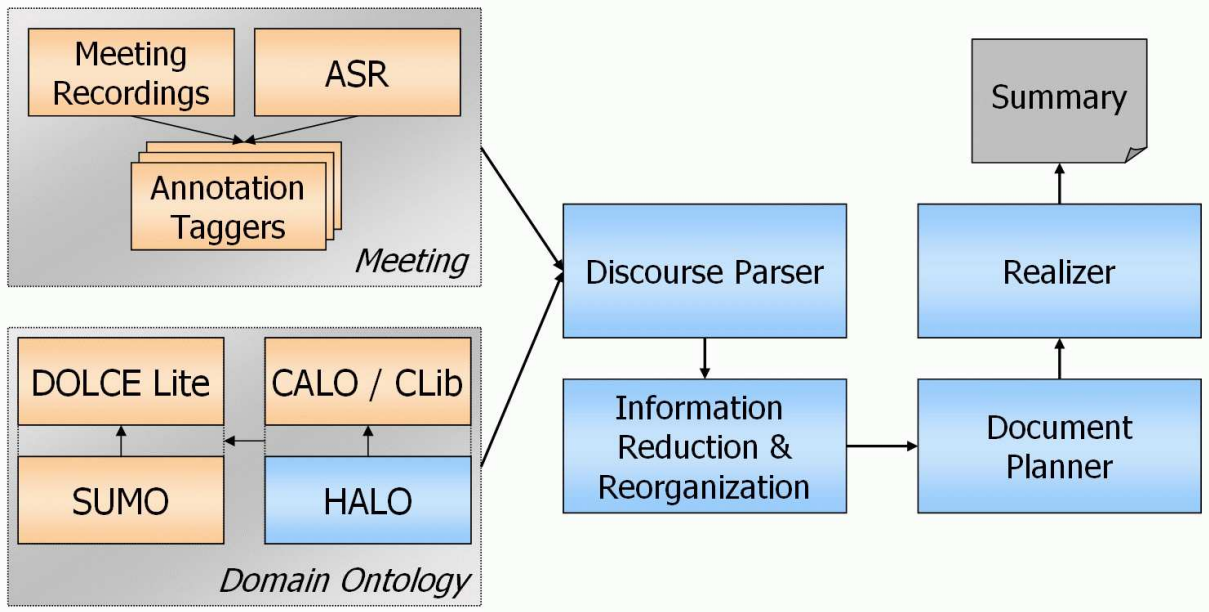

Figure 19: Abstract Architecture of ABSURD

[22] For reasons of simplicity, we assume that the source document is a piece of written text; the outlined approach, however, is not limited to a certain modality or combination of modalities.

15.2 Architecture

The design integrates components developed at DFKI (blue boxes) as well as work by other AMI partners (orange boxes)[23]. The common representation data structure of *meaning* is a combination of propositional content (see sectoin 7) and dialog acts (see section 2) and therefore faces the limitations that come with both annotation schemes, most notably the domain dependence of the propositional content ontology.

Based on the HALO ontology currently being developed at DFKI, the document parser contains a set of transformation rules to analyze the discourse items of a meeting. For the latter, the speech transcripts (either as ASR output or as hand annotation) are passed to the discourse parser component in NXT format ([28]). For coherence of the internal representation of the discourse analysis, the dialog act segment boundaries as derived by the DA annotation module are also used as atomic units for propositional content annotation. Through application of the transformation rules, each discourse segment is analyzed and an ontological representation of its propositional content is created.

Future work will concentrate on methods to tie these context-free atomic units of propositional content together to form higher level meeting structures. Here, we are aiming to extend and port the work of [6] to multiparty settings, resulting in structures similar to the game/move graphs in [121]. The result would be a holistic representations of the propositional content of the *whole* meeting.

This data structure is then passed to the "information reduction & reorganization" component. The terms reduction and reorganization stand for two different transformation techniques, both of which target the condensing of the informational graph structure. Reduction techniques attempt to downsize the graph by cutting off subgraphs that are considered to contain no relevant information. Typically, reduction would be applied to remove the representations of utterances like "uh", "yeah", etc. This has to be done with care, however, because dependent on the focus of summarization, even filler sounds may carry important information. A "yeah" sound, for instance, could have been uttered by the speaker as an explicit acknowledgment of what his predecessor has said. If the focus of the summary is to list argumentative structures, such information is not supposed to be missed. However, discourse extracts like

Speaker-1: Yeah.

Speaker-2: Yeah.

Speaker-3: Yeah.

could be combined to *one* instance of an ontology concept such as "GroupApproval". In that case, information would not be dropped, it would rather be aggregated. This technique is expressed by the term "information reorganization". Here, the internal graph representation is reduced, but without information loss.

The decision, in which case one of the reduction or reorganization techniques can be applied to a certain subgraph, can only be made on the basis of a given relevance measure. Such a measure can be seen as a parameter to the component, and allows for different "views" of summarization. For example, the industrial designer might consider only material related issues relevant while the marketing expert cares for sales and marketing topics. Therefore, appropriate summaries for these two would differ considerable. ABSURD can cope with different views or interests when provided with different relevance measures for each type of summary. The resulting structure of the condensed propositional content representation depends on the underlying relevance measure.

To deliver the information from the internal representation, the "document planner" component works in two logical steps. First, it prepares the structure of the document together with instructions on how the informational content of the summary's internal representation has to be realized. These instructions together with the prestructured summary skeleton are then passed to ABSURD's "realizer" component where the information units are finally transformed into surface forms, i.e. words and sentences and in the future also multimedia links and objects. In this process, the Realizer not only asserts that the resulting text contains all required information, it is also responsible for a coherent text flow and high readability.

[23]The CALO/CLib ontology is being developed in the CALO project `http://www.ai.sri.com/project/CALO`

15.3 Outlook

One long-term goal of the project is to enable ABSURD to generate different summaries for different target platforms. Although it may not necessarily be a near-term aspect, this consideration is already integrated in the architecture. Through the modular pipeline each supported output facility would get a particularly tailored document planner implementation. For instance, the system could manage one document planner for hypertext output, one for an electronic slide format, one for a PDA display, and so on. This means that the target platform would become an input parameter for ABSURD which in turn would select the appropriate document planner in its pipeline at run-time to generate the best suited summary.

16 Automatic Video Editing

The goal of an automatic video editing algorithm is to select relevant information from multiple video sources and present this information in a way that is ”pleasant” to human observer. This means that only picture of one camera or blended picture of several cameras is chosen and shown at the each moment of the meeting. It is also supposed that it is not necessary to replay the whole duration of the meeting - some kind of summarization can be provided by the algorithm. However, primary idea is to create output videos that will satisfy qualitative requirements of the viewers. Some elementary knowledge of film or TV production should be respected. The following criteria have to be respected for satisfaction of quality of the output videos:

- technical aspects
- aesthetical aspects
- explicit user requirements

Satisfaction of first mentioned aspects ensure that produced video contains as much of the relevant information as possible, for example following the talking participant. The aesthetical aspects ensure that particular shots are organized in a suitable form e.g. too long or too short shots are eliminated. Other rules defining how shots can be combined are known in film theory. Last aspects represent specific requirements of the viewers. For example, the viewer can prefer certain meeting participant or an activity.

Generated videos should also satisfy some structural features. Similar kind of programmes can have common structure e.g. the same parts and the order of the parts. A skeleton of given programme type can be defined and then used for generation of output videos. Specific aspects can be preferred in different parts of such model. Fig. 20) shows simple skeleton of the programme that contains meeting.

The proposed video editing algorithm is based on evaluation of activities, which occur in the meeting room. A simulation of human editor is included for the selection of the best camera and effects in given time point. Currently, the features describing physical activity of meeting participants are evaluated from detected head and hands positions [1]. Other participants' activities are deduced from the speaker identification. The problem of camera selection at given frame t from the recording with length l can be defined as discrete function:

$$c(t) = f(t, \vec{a}_1, ..., \vec{a}_n, \vec{s}) \tag{12}$$

The result of this function determines which camera should be displayed. Measured activities are presented as vectors $\vec{a}_i$ (e.g. speech, gestures, ...). Particular elements of such vector represent values of one source feature (activity) in the appropriate time points.

$$\vec{a}_i = (a_i(0), ..., a_i(m)) \tag{13}$$

The camera selection function contains in addition a state vector $\vec{s}$. It is clear that result of the camera selection function depends on previous steps of the evaluation e.g. history of the selected cameras can be stored in this vector.

$$\vec{s} = (s_1, ..., s_k) \tag{14}$$

It is supposed that camera selection process will be applied sequentially from the beginning to the end of recorded data. The state vector can be modified in every step of this evaluation. The video editing algorithm can be used in two basic applications. If activities' vectors contain data available only till time t ($m = t$), the output can be generated on the fly during the recording process, so that the meeting can be broadcasted live. On the other hand, the offline production of the output videos can use vectors compounded from activities computed during whole time period of the meetings ($m = l$). Better visual results can be achieved in the offline editing, because the camera can be also selected according to the events, which offer after the evaluated time point.

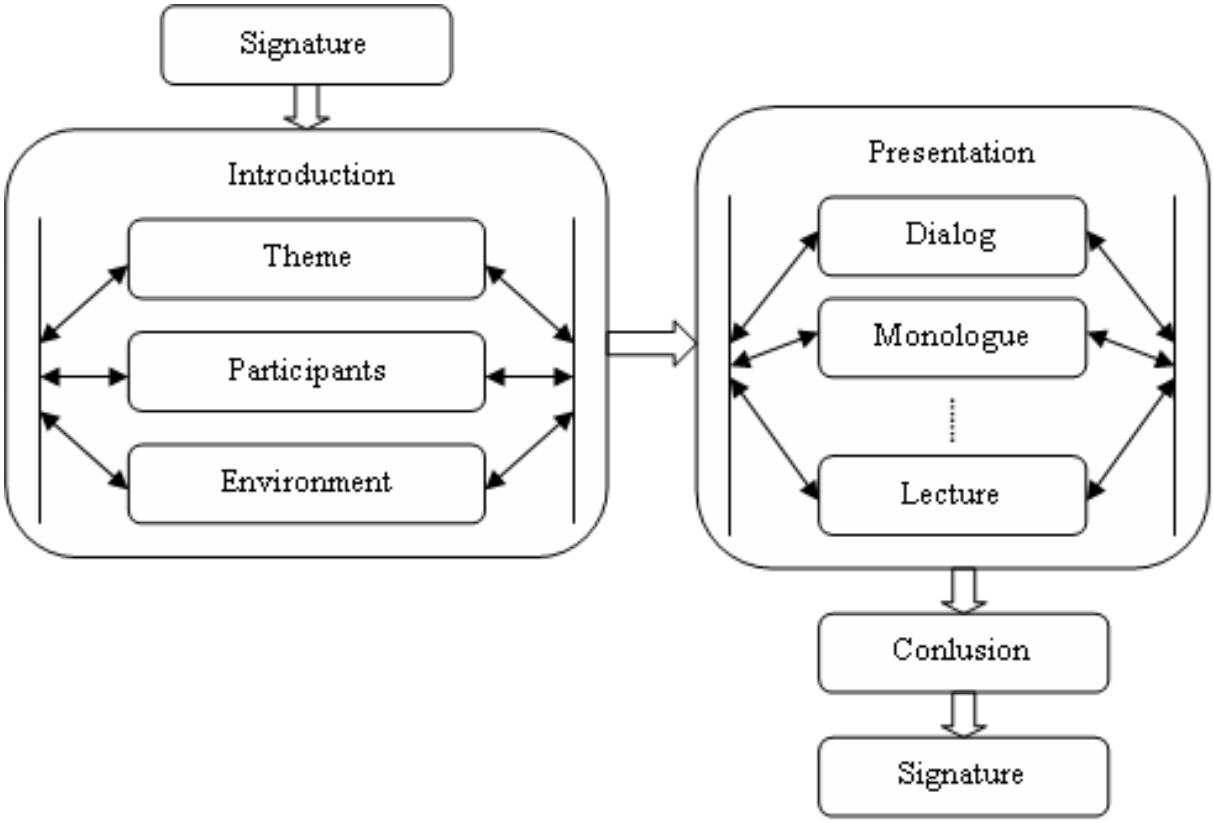

Figure 20: Example of programme skeleton

The camera selection function and the rest of the video editing algorithm are currently implemented using various rules [2]. Some of these rules describe how to convert source features into data expressing importance of an appropriate activity. Other rules represent the video editing methodology that says which cameras and when should be selected? This means that the camera showing the most important events is selected, but also the measure of desirability of the continuation of given shots is taken into account. The result of each rule is a number representing weight of one aspect e.g. activity of one participant or importance of given camera selection according to video editing methodology. All rules are connected into the network. The total weights describing importance of every camera are computed and the camera with the highest weight is selected.

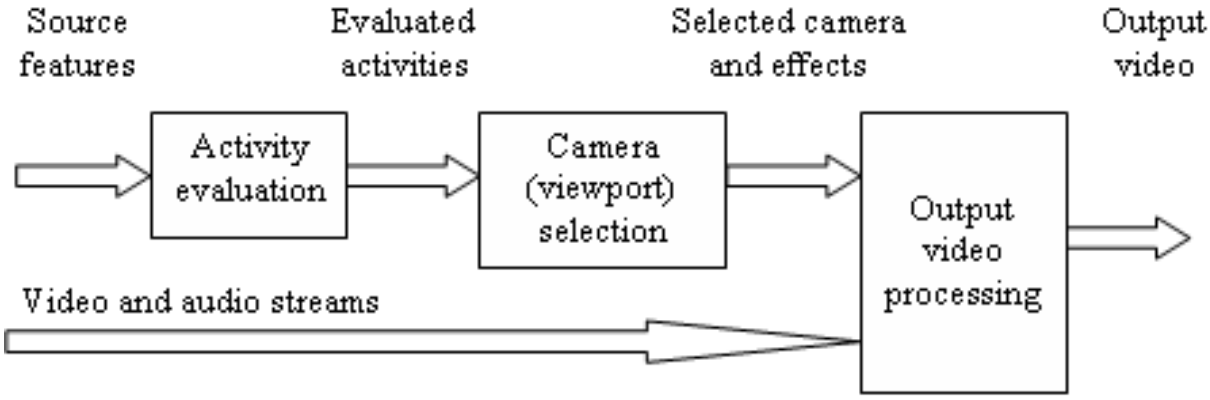

Figure 21: Video editing system

In addition to physical cameras, the system works also with so called virtual cameras. Virtual cameras are defined by position and zoom of selected part of source video. They can be used for example if a detail shot of certain participant is necessary but only distant shot with this participant is available. This is for example the case of omni-directional system where all participants are visible in the view of one common camera. Some viewports displaying particular participants or neighboring participant can be represented with virtual cameras and video editing algorithm will be working with this cameras instead of picture of physical camera [3].

The whole video editing system can produce generic full length cut of the recorded meeting offline. It is also possible to create video according to specific query given by the viewer. The selected participants or activities can be highlighted due to strengthening of an appropriate rule weight. Further, it is possible to produce a shortened version of the meeting. A summarization method based on skipping of segments with low importance is applied. In addition, video editing algorithm can be used for live broadcasting of the meetings in real-time.

References

[1] Potucek, I., Sumec, S., Spanel, M.: *Participant activity detection by hands and face movement tracking in the meeting room*, In: 2004 Computer Graphics International (CGI 2004), Los Alamitos, US, IEEE CS, 2004, s. 632-635, ISBN 0-7695-2717-1.

[2] Sumec, S.: *Multi-Camera Automatic Video Editing*, In: Proceedings of ICCVG 2004, Warsaw, Poland, 2004.

[3] Sumec, S., Potucek, I., Zemcik, P.: *Automatic Mobile Meeting Room*, In: Proceedings of 3IA'2005 International Conference in Computer Graphics and Artificial Intelligence, Limoges, FR, 2005, s. 171-177, ISBN 2-914256-07-8.

A Transcript AMI-FOBM6

P2: Ants are the most intelligent animals in the world.

P0: Well taken as a whole maybe, but individually no

P2: ?? cats

P3: Yeah but there's an 'S' VOC laugh. There is a problem here

P0: Well it's a species, a species yeah

P3: I would say the most intelligent animal is in singular

P2: Which one?

P3: Or maybe we have to consider we have to consider intelligence as a group maybe?

P0: a ?? a cat, a cow or ??

P0: Um

P3: Cause cow as a group, I would bet on cow VOC laugh. I think we we can eliminate cow anyway

P3: It doesn't look very intelligent. You have any clue of vow intelligence? VOC laugh

P0: I think they have some kind of secret manifestation of intelligence.

P2: Oh yes no no no no P0: They hide it very well. Well you can't because when they're observed, they instantly hide it. So you can't know. P3: When when they

P0: This was a guess I think.

P2: So the mother um I would rate it as ants cats, ants cats and cows

P0: What?

P2: In that in that order I'd rate them as ?? VOC laugh

P3: I would rate cats, cow, ants

P0: I would say ants.

P2: Ants yeah

P0: Yeah

P3: You would say ants first

P0: As a group

P2: Yeah as a group yeah

P3: As a group yeah but that's not really intelligence that's organization

P2: Well

P1: Um yeah yes um as an organization they are very intelligent

P2: Um the cats hardly live together, you know

P3: Yeah but is-t it can be a proof of intelligence if they can um they can have um critique opinion against other cats, where as ants just agree, so they don't really

P0: Yeah

P0: What doe-s what does it prove? is it just

P1: Actually an interesting point is that ants have survise/survived o-n on the earth for millions of years without evolution

P0: Well they have a very plastic if it's English, plastic nature. They can be modified at will the the the quee-n

P1: They can't

P3: They can

P0: The queen decides what she produces depending on the conditions

P2: That is Bees right?

P0: No I think it's true for ants

P2: Ants also

P0: So

P3: All- all of this it true, but it this not related to intelligence. Yeah good a good adaptation capacity they have good group behaviour, but they don't have any initiative or

P0: Well yeah but

P1: What's intelligence?

P1: What I'm trying to say

P2: Well cats have initiative to steal food for themselves

P3: Yeah if you let something anywhere a cat will try to

P2: Ants

P2: Ants do have the same instinct you leave your sugar box open anywhere they come there and they make it you know VOC laugh VOC laugh VOC laugh

P0: Yeah

P1: I-f if there's something, an ant will eventually find it

P3: That's much more difficult with a cow. VOC laugh If you leave something in a kitchen, you are less likely to find a cow VOC laugh VOC laugh u-

P0: You know you are in trouble yeah?

P1: It depends if the cow is very hungry. P2: Well cow usually, well cows usually, well I don't know here, but in India the cows usually have a tendency to go into an some others field to eat the green grass if it doesn't gets it.

Well depending on the situation the cow can also become intelligent

P0: ?? a mad cow maybe VOC laugh

P3: Ok

P3: Um yeah P2: Well it once like

P0: I don't know I see but ants built, they're able to built um well they modify our gardens

P2: Yeah

P0: Cats can't

P0: Yeah ants can built big structure, very complex things

P3: Yeah

P0: High span and

P1: What do you mean by modifying the environment? If you put a cat in an environment with a a lot of rats

P1: It will change the ??

P0: Yeah it is not really building

P2: So

P1: So we are still divided I think

P3: I think um that that's strange too because intelligence as a group group intelligence

P0: Yeah

P0: Well if you look at the brain

P3: Yeah

P0: We could look at it this way

P3: Yeah yeah but that is different individual yeah that can yeah that's interesting

P0: I don't think ?? I wouldn't look at an ant as a brilliant individual of I mean by itself it's nothing right?

P3: Yeah

P3: Yeah ok

P2: Ah well it is well you should look for that um story of other than French

P0: Yeah yeah you're right

P0: Yeah same

P3: I vote for ant as well

P1: Me too

References

[1] Steve Abney. Parsing by chunks. In Robert C. Berwick, Steven P. Abney, and Carol Tenny, editors, *Principle-Based Parsing: Computation and Psycholinguistics*, pages 257–278. Kluwer Academic Publishers, Boston, 1991. http://citeseer.ist.psu.edu/abney91parsing.html.

[2] Steve Abney. Partial parsing via finite-state cascade. *Journal of Natural Language Engineering*, 2(4): 337–344, 1996. http://www.vinartus.net/spa/97a.ps.

[3] M. Al-Hames, A. Dielmann, D. Gatica-Perez, S. Reiter, S. Renals G. Rigoll, and D. Zhang. Multimodal integration for meeting group action segmentation and recognition. *To Appear in Proc. Multimodal Interaction and Related Machine Learning Algorithms Workshop (MLMI–05)*, 2005.

[4] M. Al-Hames and G. Rigoll. A multi-modal graphical model for robust recognition of group actions in meetings from disturbed videos. In *Proc. IEEE ICIP*, Italy, 2005.

[5] M. Al-Hames and G. Rigoll. A multi-modal mixed-state dynamic Bayesian network for robust meeting event recognition from disturbed data. In *Proc. IEEE ICME*, 2005.

[6] Jan Alexandersson. *Hybrid Discourse Modelling and Summarization for a Speech-to-Speech Translation System*. PhD thesis, Universtit"at des Saarlandes, Germany, 2003. http://www.dfki.de/ janal/public_archive/thesis.pdf.

[7] J. Allan, J.G. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[8] Michael Argyle. *Social Interaction*. Tavistock Publications, 1973.

[9] B. Arons. Pitch-based emphasis detection for segmenting speech recording. In *Proceedings of International Conference on Spoken Language Processing*, volume 4, pages 1931–1934, 1994.

[10] Nicholas Asher and Alex Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.

[11] I. Bakx, K. van Turnhout, and J. Terken. Facial orientation during multi-party interaction with information kiosks. In *Proceedings of the Interact 2003*, 2003.

[12] R.F. Bales. *Interaction Process Analysis*. Addison-Wesley, 1951.

[13] R.F. Bales and S.P. Cohen. *SYMLOG: A System for the Multiple Level Observation of Groups.* The Free Press, 1979.

[14] M.C. Beardsley. *Practical Logic*. Prentice Hall, 1950.

[15] D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Machine Learning*, 34: 177–210, 1999.

[16] S. Bengio. An asynchronous hidden markov model for audio-visual speech recognition. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in NIPS 15*. MIT Press, 2003.

[17] J. Bilmes. Graphical models and automatic speech recognition. *Mathematical Foundations of Speech and Language Processing*, 2003.

[18] D. M. Blei and P. J. Moreno. Topic segmentation with an aspect hidden Markov model. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval*. ACM Press, 2001.

[19] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Marmuthh. Occam's razor. In *Information Processing Letters*, pages 377–380. 24 edition, 1987.

[20] Mila Boldareva. An architecture for interactive retrieval from meeting recordings. In *Proceedings of MLMI'05*, Edingburgh, UK, July 2005.

[21] R. Briggs and G. Vreede. Thinklets: Achieving predictable, repeatable, patterns of group interaction with group support systems (gss). In *Proceedings of the 34th Hawaii International Conference on System Sciences*, 2001.

[22] J.M. van Bruggen. *The use of external representations of argumentation in collaborative problem solving*. PhD thesis, Open Universiteit Nederland, June 2003.

[23] S. Buckingham Shum. Negotiating the construction and reconstruction of organisational memories. *Journal of Universal Computer Science*, 3(8):899–??, 1997.

[24] S. Buckingham Shum. *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, chapter The Roots of Computer Supported Argument Visualization. Springer Verlag, London, UK., 2003.

[25] Anne Hendrik Buist, Wessel Kraaij, and Stephan Raaijmakers. Feasibility study of extractive meeting summarization. In *CLIN 2004, The 15th Meeting of Computational Linguistics in the Netherlands*, Leiden, The Netherlands, December 2004.

[26] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. ACM SIGIR*, pages 335–336, 1998.

[27] 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 meeting corpus: A pre-announcement. *To Appear in Proc. Multimodal Interaction and Related Machine Learning Algorithms Workshop (MLMI–05)*, 2005.

[28] J. Carletta, S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voormann. The NITE XML toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers*, 35(3): 353–363, 2003.

[29] Nitesh V. Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced datasets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6, 2004.

[30] H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals. Maximum entropy segmentation of broadcast news. In *Proceedings of the IEEE International Conference on Acoustic, Speech, and Signal Processing*, Philadelphia, USA, 2005.

[31] Herbert H. Clark and B. Carlson, Thomas. Hearers and speech acts. In *Arenas of Language Use (H.H.Clark ed.)*, pages 205–247. University of Chicago Press and CSLI, 1992.

[32] J. Conklin and M.L. Begeman. gibis: a hypertext tool for exploratory policy discussion. *ACM Trans. Inf. Syst.*, 6(4):303–331, 1988. ISSN 1046-8188.

[33] G. Cooper and E. Herskovits. Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[34] T.H. Cormen, C.E. Leiserson, and R.L. Rivest. *Introduction to Algorithms*. MIT Press, 1990.

[35] S. Dharanipragada, M. Franz, J.S. McCarley, K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. Statistical methods for topic segmentation. In *ICSLP-2000*, pages 516–519, 2000.

[36] R. Dhillon, S. Bhagat, H Carvey, and E. Shriberg. Meeting recorder project: Dialogue act labeling guide. Technical report, ICSI Speech Group, Berkeley, USA, 2003. URL `http://www.icsi.berkeley.edu/Speech/mr/`.

[37] A. Dielmann and S. Renals. Multistream dynamic Bayesian network for meeting segmentation. *Lecture Notes in Computer Science*, 3361:76–86, 2005.

[38] S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141–151, September 2000.

[39] J. A. Edwards. *Handbook of Discourse*, chapter Transcription in Discourse, pages 321–348. Mass: Blackwell Publishers, 2001.

[40] F.H. Eemeren. A glance behind the scenes: The state of the art in the study of argumentation. *Studies in Communication Sciences*, 3(1):1–23, 2003.

[41] C.(S.) Ellis and P. Barthelmess. The Neem dream. In *Proceedings of the 2003 conference on Diversity in computing*, pages 23–29. ACM Press, 2003. ISBN 1-58113-790-7.

[42] Ellyson and Dovidio. *Power, Dominance, an Nonverbal Behavior*. Springer Verlag, 1985.

[43] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. Discourse segmentation of multi-party conversation. In *Proceedings of the ACL*, 2003.

[44] M. Gavalda, K. Zechner, and G. Aist. High performance segmentation of spontaneous speech using part of speech and trigger word information. In *Proceedings of the Fifth ANLP Conference*, pages 12–15, 1997.

[45] E. Goffman. Replies and responses. *Language in Society*, 5:257–313, 1976.

[46] Erving Goffman. On face-work. In *Interaction Ritual: Essays on Face to Face Behavior*, pages 5–45. New York: Doubleday Anchor, 1967.

[47] Erving Goffman. Footing. In *Forms of Talk*, pages 124–159. University of Pennsylvania Press, 1981.

[48] Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In *Proc. ACM SIGIR*, pages 19–25, 2001.

[49] T. Govier. *A practical study of argument, Sixth edition*. Wadsworth Publishing, 2005.

[50] H.P. Grice. *Logic and conversation*, chapter Syntax and Semantics: Speech Acts, pages 41–58. Academic Press, 1975.

[51] T. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

[52] Nicola Guarino. Formal ontology, conceptual analysis and knowledge representation. *Int. J. Hum.-Comput. Stud.*, 43(5-6):625–640, 1995. ISSN 1071-5819.

[53] T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, I.Mc.Cowan, J.Vepa, and S.Renals. An investigation into transcription of conference room meetings. *Submitted to Eurospeech*, 2005.

[54] M. Hearst. Texttiling: Segmenting text into multiparagraph subtopic passages. *Computational Linguistics*, 25(3):527–571, 1997.

[55] D. Hellriegel, J.W. Slocum Jr., and R.W. Woodman. *Organizational Behavior, seventh edition*. West publising company, 1995.

[56] L.R. Hoffmann. Applying experimental research on group problem solving to organizations. *Journal of applied behavioral science*, 15:375–391, 1979.

[57] E. Husserl. *Formalontologie, Part 1, Form und Wesen*. Niemeyer, 1965.

[58] Roman Ingarden. *Der Streit um die Existenz der Welt. Vol. 3: Über die kausale Struktur der realen Welt*. Max Niemeyer, Tübingen, Germany, 1974.

[59] N. Jovanovic, R. Op den Akker, and A. Nijholt. A corpus for studying addressing behavior in multi-party dialogues. In *Proc. of The sixth SigDial conference on Discourse and Dialogue*, 2005. Submitted.

[60] M. Kan. *Automatic text summarization as applied to information retrieval: Using indicative and informative summaries*. PhD thesis, Columbia University, New York USA, 2003.

[61] G. Kanselaar, G. Erkens, J. Andriessen, M. Prangsma, A. Veerman, and J. Jaspers. *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, chapter Desiging Argumentation Tools for Collaborative Learning. Springer Verlag, London, UK., 2003.

[62] J. Katzav, G.W.A. Rowe, and C.A. Reed. The argument research corpus. In *Working notes of the conference on Practical Applications of Liguistic Corpora (PALC)*, pages 77–83, 2003.

[63] M. Katzenmaier, R. Stiefelhagen, and T. Schultz. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *ICMI 2004*, October 2004.

[64] A. Kendon. Some functions of gaze direction in social interaction. *Acta Psychologica*, 32:1–25, 1967.

[65] J.L. Kestler. *Questioning Techniques and Tactics*. McGraw-Hill, 1982.

[66] W. Kunz and H.W.J. Rittel. Issues as elements of information systems. Working Paper WP-131, Univ. Stuttgart, Inst. Fuer Grundlagen der Planung, 1970.

[67] J. Kupiec, J. Pederson, and F. Chen. A trainable document summarizer. In *ACM SIGIR '95*, pages 68–73, 1995.

[68] Mirella Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *ACL*, pages 545–552, 2003.

[69] G. Lathoud, I. A. McCowan, and J.-M. Odobez. Unsupervised Location-Based Segmentation of Multi-Party Speech. In *Proc. 2004 ICASSP-NIST Meeting Recognition Workshop*, 2004.

[70] C.-Y. Lin and E. H. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. HLT-NAACL*, 2003.

[71] A. Lisowska. Multimodal interface design for the multimodal meeting domain: Preliminary indications from a query analysis study. Technical report, ISSCO/TIM/ETI, Universit de Genve, Switserland, november 2003. IM2.MDM Report 11.

[72] Y. Liu, E. Shriberg, A. Stolcke, and M. Harper. Using machine learning to cope with imbalanced classes in natural sppech: Evidence from sentence boundary and disfluency detection. In *Proceedings of the Intl. Conf. Spoken Language Processing*, 2004.

[73] Inderjeet Mani, Barbara Gates, and Eric Bloedorn. Improving summaries by revising them. In *Proceedings of the 37th conference on Association for Computational Linguistics*, pages 558–565, Morristown, NJ, USA, 1999. Association for Computational Linguistics. ISBN 1-55860-609-3.

[74] W.C. Mann and S.A. Thompson. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, University of Southern California, 1987.

[75] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(3):305–317, 2005.

[76] J. Moore, M. Kronenthal, and S. Ashby. Guidelines for AMI speech transcriptions. Technical report, IDIAP, Univ. of Edinburgh, February 2005.

[77] N. Morgan, D. Baron, S. Bhagat, H. Carvey, R. Dhillon, J. Edwards, D. Gelbart, A. Janin, A. Krupski, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. Meetings about meetings: research at icsi on speech in multiparty conversations. *Proc. IEEE ICASSP*, 2003.

[78] G. Murray, S. Renals, and J. Carletta. Extractive summarization of meeting recordings. 2005.

[79] A. Neass. *Communication and argument. Elements of applied semantics.* George Allen & Uwin Press, 1966.

[80] S.E. Newman and C.C. Marshall. Pushing toulmin too far: Learning from an argument representation scheme. Technical Report SSL-92-45, Xerox PARC, 1991.

[81] Y. Ohsawa, N. Matsumura, and M. Ishizuka. Influence diffusion model in text-based communication. In *Proc. of The eleventh world wide web conference*, 2002. ISBN 1-880672-20-0.

[82] N. Oliver, E. Horvitz, and A. Garg. Layered representations for learning and inferring office activity from multiple sensory channels. In *Proc. ICMI*, Pittsburgh, Oct. 2002.

[83] Miles Osborne. Shallow parsing as part-of-speech tagging. In Claire Cardie, Walter Daelemans, Claire Nedellec, and Erik Tjong Kim Sang, editors, *Proceedings of CoNLL-2000 and LLL-2000*, pages 145–147, Lisbon, Portugal, 2000. `http://acl.ldc.upenn.edu/W/W00/W00-0731.pdf`.

[84] Miles Osborne. Shallow parsing using noisy and non-stationary training material. *Journal of Machine Learning Research*, 2(4):695–718, 2002. `http://www.jmlr.org/papers/volume2/osborne02a/osborne02a.pdf`.

[85] V. Pallotta, J. Niekrasz, and M. Purver. Collaborative and argumentative models of meeting discussions. In *Proceeding of CMNA-05 international workshop on Computational Models of Natural Arguments (part of IJCAI 2005)*, July 2005.

[86] V. Pavlovic, B. Frey, and T.S. Huang. Time series classification using mixed-state dynamic Bayesian networks. In *Proc. IEEE CVPR*, 1999.

[87] L. Pevzner and M. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, 2002.

[88] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE*, 2(77):257–286, 1989.

[89] Lance Ramshaw and Mitch Marcus. Text chunking using transformation-based learning. In David Yarovsky and Kenneth Church, editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94, 1995. `http://acl.ldc.upenn.edu/W/W95/W95-0107.pdf`.

[90] Adwait Ratnaparkhi. A maximum entropy part-of-speech tagger. In Eric Brill and Kenneth Church, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing 1996*, pages 133–142, 1996. `http://acl.ldc.upenn.edu/W/W96/W96-0213.pdf`.

[91] C. Reed and G. Rowe. Araucaria: Software for puzzles in argument diagramming and xml. Technical report, Department of Applied Computing, University of Dundee, 2001.

[92] D. Reidsma, R. op den Akker, Rienks R., Poppe R., A. Nijholt, D. Heylen, and J. Zwiers. Virtual meeting rooms: From observation to simulation. In *Proc. of the Social Intelligence Design Workshop*, 2005. to appear.

[93] D. Reidsma, R. Rienks, and N. Jovanovic. Meeting modelling in the context of multimodal research. In *Proc. of the Workshop on Machine Learning and Multimodal Interaction*, 2004.

[94] S. Reiter and G. Rigoll. Segmentation and classification of meeting events using multiple classifier fusion and dynamic programming. In *Proc. IEEE ICPR*, pages 434–437, 2004.

[95] S. Reiter and G. Rigoll. Multimodal meeting analysis by segmentation and classification of meeting events based on a higher level semantic approach. In *Proc. IEEE ICASSP*, 2005.

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

[97] R. Rienks, A. Nijholt, and D. Reidsma. *Meetings and Meeting support in ambient intelligence*, chapter In Ambient Intelligence, Wireless Networking, Ubiquitous Computing. Artech House, Norwood, MA, USA, 2005. In Press.

[98] D. Schum and A. Martin. Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review*, 17(1):105–152, 1982.

[99] R. Schwartz. *The skilled facilitator*. John-Bass Publishers, 1994.

[100] A. Selvin, S. Buckingham Shum, M. Sierhuis, J. Conklin, B. Zimmermann, C. Palus, W. Drath, D. Horth, J. Domingue, E. Motta, and G. Li. Compendium: Making meetings into knowledge events. In *Proc. Knowledge Technologies 2001*, March 2001.

[101] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, , and H. Carvey. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proc. 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100, 2004.

[102] E. Shriberg, A. Stolcke, D. Hakkani-tur, and G. Tur. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communications*, 31(1-2):127–254, 2000.

[103] Weltym C. Smith, B. Fois introduction: Ontology—towards a new synthesis. In *FOIS '01*, pages 3–9, New York, NY, USA, 2001. ACM Press. ISBN 1-58113-377-4.

[104] P. Smolensky, B. Fox, R. King, and C. Lewis. *Computer-aided reasoned discourse, or, how to argue with a computer*, pages 109–162. Lawrence Erlbaum, 1988.

[105] J. Steinberger and K. Ježek. Using latent semantic analysis in text summarization and summary evaluation. In *Proc. ISIM '04*, pages 93–100, 2004.

[106] R. Stiefelhagen and J. Zhu. Head orientation and gaze direction in meetings. In *Conference on Human Factors in Computing Systems (CHI2002)*, April 2002.

[107] N. Stokes, J. Carthy, and A.F. Smeaton. Select: a lexical cohesion based news story segmentation system. *AI Communications*, 17(1):3–12, January 2004.

[108] D. Suthers, A. Weiner, J. Connelly, and M. Paolucci. Belvedere: Engaging students in critical discussion of science and public policy issues. In *Proceedings of the the 7th World Conference on Artificial Intelligence in Education (AIED).*, August 1995.

[109] D.D. Suthers. Towards a systematic study of representational guidance for collaborative learning discourse. *Journal of Universal Computer Science*, 7(3), 2001. Electronic Publication.

[110] Erik F. Tjong Kim Sang and Sabine Buchholz. Introduction to the conll-2000 shared task: Chunking. In Claire Cardie, Walter Daelemans, Claire Nedellec, and Erik Tjong Kim Sang, editors, *Proceedings of CoNLL-2000 and LLL-2000*, pages 127–132. Lisbon, Portugal, 2000. `http://acl.ldc.upenn.edu/W/W00/W00-0726.pdf`.

[111] S. Toulmin. *The Uses of Argument*. Cambridge University Press, 1958.

[112] Dolf Trieschnigg and Wessel Kraaij. Hierarchical topic detection in large digital news archives, exploring a sample based approach. In *Proceedings of the Fifth Dutch-Belgian Information Retrieval Workshop*, pages 55–62, Utrecht, The Netherlands, January 2005. Utrecht University.

[113] A. Tritschler and R.A. Gopinath. Improved speaker segmentation and segments clustering using the bayesian information criterion. In *Proc. EUROSPEECH '99*, 1999.

[114] R. Valenza, T. Robinson, M. Hickey, and R. Tucker. Summarization of spoken audio through information extraction. In *Proc. ESCA Workshop on Accessing Information in Spoken Audio*, pages 111–116, 1999.

[115] F. Van Eemeren, R. Grootendorst, and F. Snoeck Henkemans. *Argumentation*. Lawrence Erlbaum Associates, 2002.

[116] T. Van Gelder. *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, chapter Enhancing Deliberation through Computer-Supported Argument Visualization. Springer Verlag, London, UK., 2003.

[117] T.J. van Gelder. Argument mapping with reason!able. The American Philosophical Association Newsletter on Philosophy and Computers, 2002.

[118] P. van Mulbregt, J. Carp, L. Gillick, S. Lowe, and J. Yamron. Segmentation of automatically transcribed broadcast news text. In *Proceedings of the DARPA Broadcast News Workshop*, pages 77–80. Morgan Kaufman Publishers, 1999.

[119] A Veerman. *Computer-supported collaborative learning through argumentation*. PhD thesis, University of Utrecht, 2000.

[120] R. Vertegaal. *Look who is talking to whom*. PhD thesis, University of Twente, September 1998.

[121] Wolfgang Wahlster, editor. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, 2000.

[122] M. Walker and S. Whittaker. Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proceedings of the 28th Annual Meeting of the ACL*, 1990.

[123] D.N. Walton. *Argument Structure, A pragmatic Theory*. University of Toronto Press, 1996. ISBN 0-8020-0768-6.

[124] D.N. Walton and C.A. Reed. *Anyone Who Has a View: Theoretical Contributions to the Study of Argumentation*, chapter Diagramming, Argumentation Schemes and Critical Questions, pages 195–211. Kluwer, Dordrecht, The Netherlands, 2003.

[125] J.H. Wigmore. *The principles of Judicial Proof, 2nd ed.* Little, Brown and Company, 1931.

[126] J. Yoshimi. The structure of debate. Technical report, University of Claifornia, Merced, September 2004.

[127] K. Zechner and A. Waibel. Minimizing word error rate in textual summaries of spoken language. In *Proc. NAACL-2000*, 2000.

[128] Klaus Zechner and Alex Waibel. DIASUMM: Flexible summarization of spontaneous dialogues in unrestricted domains. In *Proceedings of COLING-2000*, pages 968–974, 2000.

[129] D. Zhang, D. Gatica-Perez, and S. Bengio. Semi-supervised meeting event recognition with adapted hmms. *Proc. IEEE Int. Conf. on Multimedia (ICME)*, July 2005.

[130] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud. Multimodal group action clustering in meetings. *Proc. ACM Int. Conf. on Multimedia, Workshop on Video Surveillance and Sensor Networks (ACM MM-VSSN)*, October 2004.

[131] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud. Modeling individual and group actions in meetings with layered hmms. In *IEEE Transactions on Multimedia, accepted for publication*, May 2005.

[132] Y. Zhang, Q. Diao, S. Huang, and W. Hu. DBN based multi-stream models for speech. *Proc. IEEE ICASSP*, 2003.

[133] M. Zobl, F. Wallhoff, and G. Rigoll. Action recognition in meeting scenarios using global motion features. In J. Ferryman, editor, *Proc. PETS-ICVS*, pages 32–36, 2003.