Towards a Decent Recognition Rate for the Automatic Classification of a Multidimensional Dialogue Act Tagset

Stephan Lesch and Thomas Kleinbauer and Jan Alexandersson[*]
DFKI GmbH
Stuhlsatzenhausweg 3, D-66123 Saarbücken
{`lesch,kleiba,janal`}`@dfki.de`

Abstract

In this paper, we present some thoughts and examinations on statistical dialogue act annotation using multidimensional dialogue act labels, based on the ICSI meeting corpus and the associated MRDA tag set. We show some statistics of this corpus, and preliminary results of a statistical tagger for the dialogue act labels, together with a proposal for a more realistic interpretation of these results.

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 categorization of *dialogue acts* [Austin, 1962; Searle, 1969] 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[1] meeting recorder project [Dhillon *et al.*, 2004], 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.

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).

[*] The research presented here is funded by the EU under the grant FP6-506811 (AMI).

[1] International Computer Science Institute at Berkeley, CA

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, 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".

The paper is organized as follows: the next section describes the MRDA tagset and a simplification thereof—the MALTUS tagset. In section 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 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 6 we conclude the paper and provide some future directions.

2 Multidimensional Tagsets

The labels of a dialog act tagset are not necessarily multidimensional. The Verbmobil System, for example [Alexandersson *et al.*, 1998], used a small set of roughly 30 tags tailored to its particular application, the automatic translation of telephone negotiations. Examples of the Vermobil tags are greet, bye, introduce, request, suggest.

Multidimensional tagsets, on the other hand, allow to annotate several aspects of an utterance. The DAMSL[2] tagset, for instance, defines four aspects: the communicative status, the information level and the forward and backward looking function of the utterance. A variant of the DAMSL tagset, the SWBD tagset [Daniel Jurafsky, 1997], was used for annotation in the Switchboard project; the SWBD tagset, in turn, served as the basis for the MRDA tagset [Popescu-Belis, 2003].

2.1 The MRDA Tagset

The "Meeting Recorder Dialogue Act" tagset was used to annotate the ICSI meeting corpus.[3] 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>(ˆ<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`).

2.2 The MALTUS Tagset

MALTUS, introduced in [Popescu-Belis, 2003], 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

(U | T1 ˆ 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).

3 Some Corpus Characteristics

The experiments presented are based on the the ICSI meeting corpus [Janin *et al.*, 2003], 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-labeled (z), and some are not labeled at all; these utterances and their successors are ignored, leaving 116097 utterances from which we take the training and testing material.

3.1 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

[2] Dialogue Act Markup on Several Layers, [Allen and Core, 1997]

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

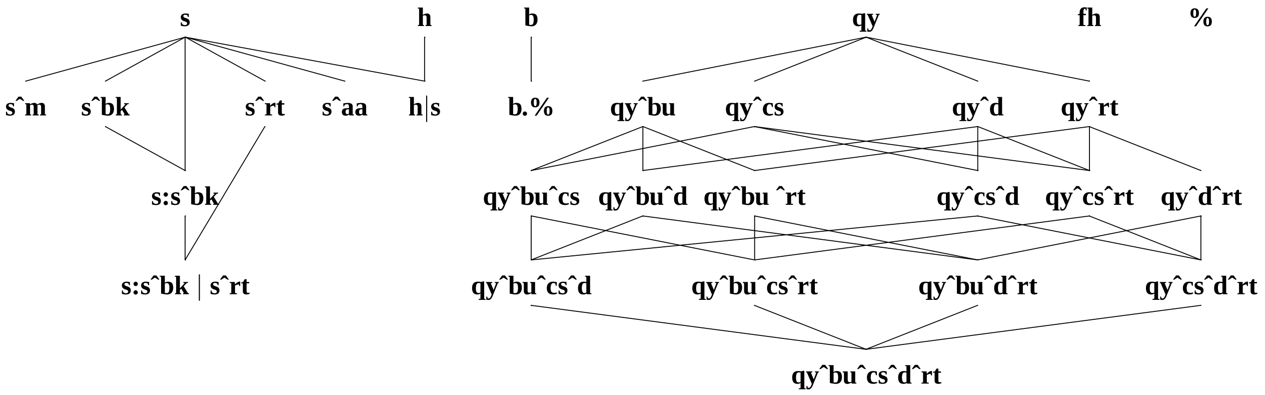

Figure 1: The lattice formed by the MRDA labels shown in table 1. Labels are ordered by the subset relation. *Compound labels*, i. e., , two labels combined with "|" or ":", are daughters of the two separate labels. Note that only the parts of the compound labels | were used in the classification experiments.

- 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.

3.2 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.

3.3 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").

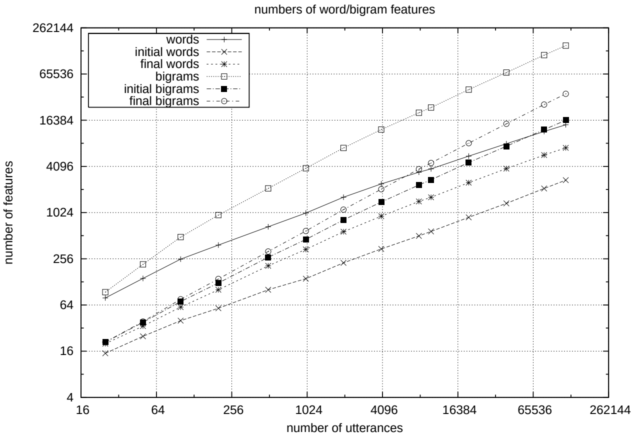

Figure 2: The number of words and bigrams for different numbers of utterances

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.

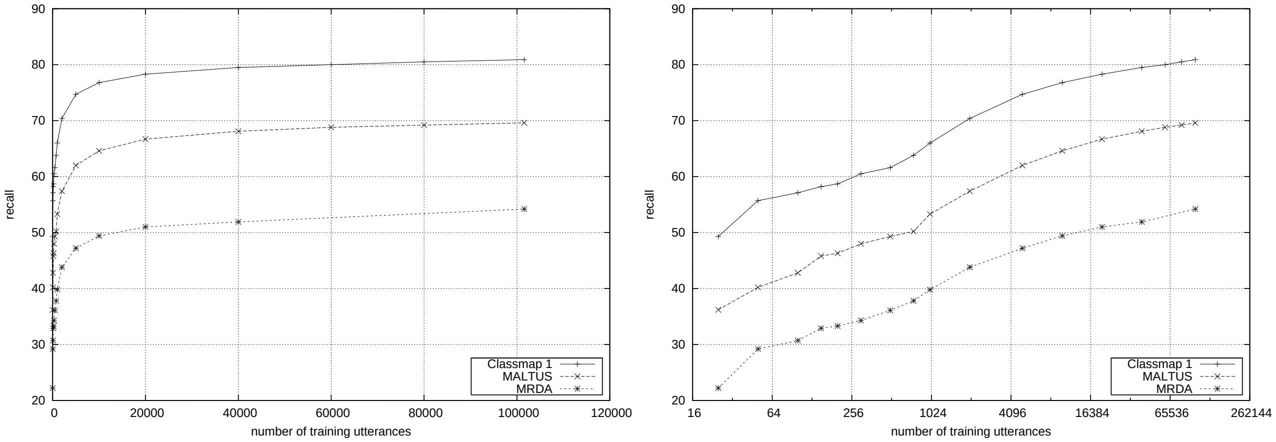

Figure 3: Recall (percent) for MRDA and MALTUS labels, and MRDA mapped with classmap 1, with different sizes of the training set. (linear and log scale, using 4-fold cross-validation, 2-fold for MRDA with 101584 training utterances)

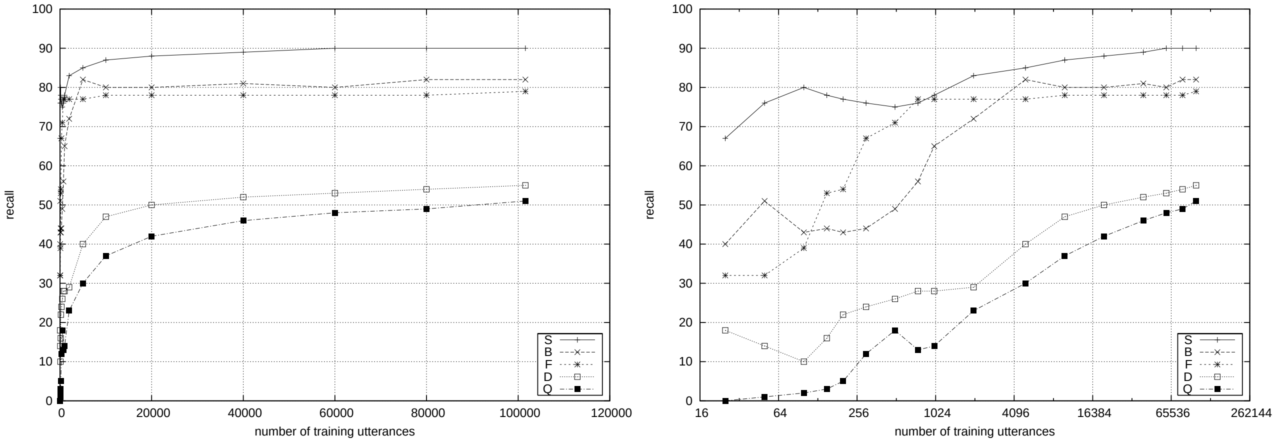

Figure 4: Recall (percent) for statements, questions, floors, backchannels and disruptions (classmap 1, linear and log scale, 4-fold cross-validation)

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 recognize 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 fefine 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 [Lesch *et al.*, 2005]). 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

$$\begin{aligned} DA^I &:= DA^T \cap DA^C && (5)\\ \delta^C &:= |DA^C| - |DA^I| && (6)\\ \delta^T &:= |DA^T| - |DA^I| && (7) \end{aligned}$$

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:

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

Note the denominators: the distances are normalized to the sizes of the true and the classified labels. Conversely, SCORRE simply normalizes 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%
total	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 dependant 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.

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.[4]

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

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

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%).

[Clark and Popescu-Belis, 2004] 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.

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

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:

- 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 [Clark and Popescu-Belis, 2004].
- 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 recognized 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—[Clark and Popescu-Belis, 2004] reports 84.9% correct classifications.

5.1 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 recognized. 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

$$H \quad := \quad -\sum_{l \in labels} p(l) log_2 p(l)$$

$$p(l) \quad := \quad \frac{number\ of\ occurrences\ of\ l}{corpus\ size}$$

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

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

Then we find the pair (m, d) in the current set which minimizes 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.

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.

#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.

6 Discussion and Outlook

We have discussed the task of dialogue act classification for a multidimensional tag set. In particular, we have focused 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., [Reithinger and Klesen, 1997; Stolcke *et al.*, 2000].

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:

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: `s^aa` (statement and accept), `s^bk` (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., [Reithinger and Klesen, 1997]. 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 [Clark and Popescu-Belis, 2004], 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 [Clark and Popescu-Belis, 2004] 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 recognizing 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 minimize an error function. This function is based on the number of incorrect classifications and does not recognize 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.

		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.

References

[Alexandersson *et al.*, 1998] 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.

[Allen and Core, 1997] 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.

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

[Clark and Popescu-Belis, 2004] 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.

[Daniel Jurafsky, 1997] 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.

[Dhillon *et al.*, 2004] 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.

[Janin *et al.*, 2003] 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.

[Lesch *et al.*, 2005] 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.

[Popescu-Belis, 2003] 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).

[Reithinger and Klesen, 1997] 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.

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

[Stolcke *et al.*, 2000] 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.