Abstract

We present recent work in the area of Dialogue Act (DA) tagging. Identifying the dialogue acts of utterances is recognised as an important step towards understanding the content and nature of what speakers say. Our experiments investigate the use of a simple dialogue act classifier based on purely intra-utterance features — principally involving word n-gram cue phrases. Such a classifier performs surprisingly well, rivalling scores obtained using far more sophisticated language modelling techniques for the corpus we address. We also discuss the potential utility of classifiers that identify the n most likely dialogue acts for each utterance, leaving it to some later process to choose amongst these alternatives.

1 Introduction

In the area of spoken language dialogue systems, the ability to assign user input with a functional tag which represents the communicative intentions behind each utterance — the utterance’s dialogue act — is acknowledged to be a useful first step in dialogue processing. Such tagging can assist the semantic interpretation of user utterances, and can help an automated system in producing an appropriate response.

Researchers, for example (Hirschberg and Litman, 1993; Grosz and Sidner, 1986), speak of cue phrases in utterances which can serve as useful indicators of dialogue acts. In common with the work of (Samuel, Carberry and Vijay-Shanker, 1999), we wanted to detect automatically word n-grams in a corpus that might serve as potentially useful cue phrases. The method we chose for selecting such phrases is based on their predictivity. The predictivity of cue phrases can be exploited directly in a simple model of dialogue act classification that employs only intra-utterance features. We report here the results of experiments evaluating this simple approach on the SWITCHBOARD corpus.Surprisingly, the results we obtain rival the best results achieved on that corpus, in work by Stolcke et al. (2000), who use a far more complex approach involving Hidden Markov modelling (HMM), that addresses both the sequencing of words within utterances and the sequencing of dialogue acts over utterances.

This simple classification approach can as well be used to produce a (possibly ranked) list of the n most likely alternative classifications for each utterance, which might feed into some subsequent process, such as a dialogue manager, that could select amongst the restricted set of alternatives offered on the basis of higher-level dialogue information. The subsequent process might alternatively be a machine-learning based component trained to make the final choice of DA based on inter-utterance context, with the possible benefit of having a much reduced feature space from the elimination of word n-gram based features, which would have already been exploited in the simple classifier component.

The work described in this paper forms part of the AMITIES project (Hardy et al., 2004), which
aims to build automated service counters allowing users to access information (e.g. such as banking information) in a more natural and flexible way. The models we use to achieve this will make use of dialogue act sequencing information.

This paper presents our work on dialogue act classification using intra-utterance information. Previous work with dialogue act modelling will be outlined in Section 2. An overview of the available corpora for this task is given in Section 3. Our experiments evaluating the simple cue-based dialogue act classifier approach are described in Section 4. Our initial explorations around classifiers assigning n-best lists of DAs is described in Section 5. We end with some discussion and an outline of intended further work.

2 Related Work

There has been an increasing interest in using machine learning techniques on problems in spoken dialogue. One thread of this work has addressed dialogue act modelling, i.e. the task of assigning an appropriate dialogue act tag to each utterance in a dialogue. It is only recently, with the availability of annotated dialogue corpora, that research in this area has become possible.

One approach that has been tried for dialogue act tagging is the use of n-gram language modelling, exploiting principally ideas drawn from the area of speech recognition. For example, (Reithinger and Klesen, 1997) have applied such an approach to the VERBMOBIL corpus, which provides only a rather limited amount of training data, and report a tagging accuracy of 74.7%. (Stolcke et al., 2000) apply a somewhat more complicated HMM method to the SWITCHBOARD corpus, one which that addresses both the sequencing of words within utterances and the sequencing of dialogue acts over utterances. They use a single split of the data for their experiments, with 198k utterances for training and 4k utterances for testing, achieving a DA tagging accuracy of 71% on word transcripts. These performance differences, with a higher tagging accuracy score for the VERBMOBIL corpus despite significantly less training data, can be seen to reflect the differential difficulty of tagging for the two corpora.

A second approach that has been applied to dialogue act modelling, by (Samuel, Carberry and Vijay-Shanker, 1998), uses transformation-based learning over a number of utterance features, including utterance length, speaker turn and the dialogue act tags of adjacent utterances. They achieved an average score of 75.12% tagging accuracy over the VERBMOBIL corpus. A significant aspect of this work, that is of particular relevance here, has addressed the automatic identification of word sequences that might serve as useful dialogue act cues. A number of statistical criteria are applied to identify potentially useful word n-grams which are then supplied to the transformation-based learning method to be treated as ‘features’.

3 Corpora

3.1 Publicly available corpora

Three key corpora have been used in most work on DA modelling. First, the VERBMOBIL project, on speech-to-speech translation, produced a corpus of 168 English annotated task-oriented dialogues, whose ‘task’ is meeting arrangement. The VERBMOBIL corpus is tagged used a total of 46 tags, which are then further clustered into 26 top-level tags. Secondly, the SWITCHBOARD

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Availability</th>
<th>Utterance count</th>
<th>Dialogue count</th>
<th>Word count</th>
<th>Distinct words</th>
<th>Dialogue type</th>
</tr>
</thead>
<tbody>
<tr>
<td>SWITCHBOARD</td>
<td>public</td>
<td>223606</td>
<td>1155</td>
<td>1431725</td>
<td>21715</td>
<td>Conversational</td>
</tr>
<tr>
<td>VERBMOBIL</td>
<td>public</td>
<td>3117</td>
<td>168</td>
<td>24980</td>
<td>959</td>
<td>Task-oriented</td>
</tr>
<tr>
<td>MAPTASK</td>
<td>public</td>
<td>26621</td>
<td>128</td>
<td>152705</td>
<td>2502</td>
<td>Task-oriented</td>
</tr>
<tr>
<td>AMITIES GE</td>
<td>restricted</td>
<td>30206</td>
<td>1000</td>
<td>228165</td>
<td>7841</td>
<td>Task-oriented</td>
</tr>
<tr>
<td>AMITIES IBM</td>
<td>restricted</td>
<td>122080</td>
<td>5000</td>
<td>1132663</td>
<td>11586</td>
<td>Task-oriented</td>
</tr>
<tr>
<td>Dialogue Act</td>
<td>% of corpus</td>
<td>Dialogue Act</td>
<td>% of corpus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------------</td>
<td>-------------------------------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>statement-non-opinion</td>
<td>36%</td>
<td>action-directive</td>
<td>0.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>acknowledge</td>
<td>19%</td>
<td>collaborative completion</td>
<td>0.4%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>statement-opinion</td>
<td>13%</td>
<td>repeat-phrase</td>
<td>0.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>agreeaccept</td>
<td>5%</td>
<td>open-question</td>
<td>0.3%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>abandoned</td>
<td>5%</td>
<td>rhetorical-questions</td>
<td>0.2%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>appreciation</td>
<td>2%</td>
<td>hold before answer</td>
<td>0.2%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes-no-question</td>
<td>2%</td>
<td>reject</td>
<td>0.2%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-verbal</td>
<td>2%</td>
<td>negative non-no answers</td>
<td>0.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes answers</td>
<td>1%</td>
<td>signal-non-understanding</td>
<td>0.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>conventional-closing</td>
<td>1%</td>
<td>other answers</td>
<td>0.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>uninterpretable</td>
<td>1%</td>
<td>conventional-opening</td>
<td>0.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wh-question</td>
<td>1%</td>
<td>or-clause</td>
<td>0.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no answers</td>
<td>1%</td>
<td>dispreferred answers</td>
<td>0.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>response acknowledgement</td>
<td>1%</td>
<td>3rd-party-talk</td>
<td>0.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hedge</td>
<td>1%</td>
<td>offers, options commits</td>
<td>0.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>declarative yes-no-question</td>
<td>1%</td>
<td>self-talk</td>
<td>0.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>other</td>
<td>1%</td>
<td>downplayer</td>
<td>0.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>backchannel in question form</td>
<td>1%</td>
<td>maybeaccept-par</td>
<td>&lt;0.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quotation</td>
<td>0.5%</td>
<td>tag-question</td>
<td>&lt;0.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>summarisereformulate</td>
<td>0.5%</td>
<td>declarative wh-question</td>
<td>&lt;0.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>affirmative non-yes answers</td>
<td>0.4%</td>
<td>apology</td>
<td>&lt;0.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2: Switchboard dialogue acts**

corpus (Jurafsky et al., 1998) comprises 1155 annotated conversations of an unstructured, non-directed character, which hence exhibit much greater semantic variability than VERBMOMIL, and have therefore been thought to present a more difficult problem for accurate DA modelling. The corpus is annotated using an elaboration of the DAMSL tag set (Core and Allen, 1997), involving 50 major classes, together with a number of diacritic marks, which combine to generate 220 distinct labels. Jurafsky et al. (1998) propose a clustering of the 220 tags into 42 larger classes, listed in Figure 2, and it is this clustered set that was used in the experiments of (Stolcke et al., 2000). The third corpus is MAP-TASK, comprising 128 task-oriented dialogues in which two people negotiate an agreed route on separate (and slightly different) maps. The DA annotation uses 12 distinct DA labels, and part of the corpus is annotated for dialogue games. The dialogues in this corpus are more collaborative in nature.

### 3.2 Restricted Corpora

Although not used in the experiments reported here, the nature of the AMITIES corpora was a motivating factor in the selection of the SWITCHBOARD corpus for our work. Previous work on the automatic selection of cue phrases from a corpus was done on VERBMOMIL, which has a very small vocabulary size and utterance count. We concentrated on SWITCHBOARD in part to study the effect of scale on this task.

During the AMITIES project, we collected 1000 English and 1000 French human-human dialogues from GE call centres. The calls are of an information seeking or transactional type, in which customers interact with their financial accounts by phone to check balances, make payments and report lost credit cards. The AMITIES GE corpus is annotated with DAS (using DAMSL) and additional domain specific semantic information such as account numbers and credit card details (Hardy et al., 2002). Later in AMITIES, we acquired 10000 transcribed calls from an IBM call centre (half French, half English), which involve call-routing dialogues, of limited length, for a product and hardware support service. This data is currently being annotated, using the same formalism as AMITIES GE. Both corpora are large (equal to SWITCHBOARD) but are task oriented and so have greater regularity in semantic content that we hope can be exploited to help DA tagging.
4 Simple DA Classification

In this section we describe our simple approach to DA classification, based on intra-utterance features, together with our experiments to evaluate it. A key aspect of the approach is the selection of the word n-grams to use as cue phrases in tagging. (Samuel, Carberry and Vijay-Shanker, 1999) investigate a series of different statistical criteria for use in automatically selecting cue phrases. We use a criterion of predictivity, described below, which is one that Samuel et al. do not consider. Predictivity values are straightforward to compute, so the approach can feasibly be applied to very large corpora. As we shall see, predictivity scores are used not only in selecting cue phrases, but also directly as part of the classification method.

4.1 Experimental corpus

For our experiments, we used the SWITCHBOARD data set of 1,155 annotated conversations. The dialogue act types for this set can be seen in (Jurafsky, Shriberg and Biasca, 1997). Altogether these 1,155 conversations comprise in the region of 205k utterances. However, we split this into three data sets. The first mirrored the size of the experiments performed on the VERBMOBIL corpus, described above, i.e. 3k utterances. To investigate the effects of using a greater amount of data, a 50k utterance data set was used, with 45k used for training and 5k for testing in each experiment. The final split used the same data size as that of (Stolcke et al., 2000), with 198k utterances for training and 4k for testing. We hoped to show that a significant increase in the amount of training data would translate to a much improved tagging accuracy. Our experiments address both the initial 220 element tag set of the corpus, and the clustered set of 42 tags discussed earlier, and listed in Figure 2. The corpus was pre-processed to remove all punctuation and case information. Some of the corpus mark-up, such as filler information described in (Meteer, 1995), was also removed.

Our experiments used a cross-validation approach, with results being averaged over 10 runs. For the first two data sets, this was a standard ten-fold approach, i.e. with the data being split into ten approximately equal partitions, each being used in turn for testing, with the remainder combined for training. Cross-validation of this kind is recognised as the standard way to estimate predictive accuracy. For the third data set, created for comparability with (Stolcke et al., 2000), the test set is much less than a tenth of the overall set, so a standard ten-fold approach does not apply. Instead, we randomly selected dialogues out of the overall data to create ten subsets of around 4k utterances for use as test sets, which were re-used across the different experimental runs. In each case, the corresponding training set was the overall data minus that subset. In addition to cross-validated results, we also report the single highest score from the ten runs performed for each experimental case. We have done this to facilitate comparison with the results of (Stolcke et al., 2000).

4.2 Cue Phrase Selection

For our experiments, the word n-grams used as cue phrases during classification are computed from the training data. All word n-grams of length 1–4 within the data are considered as candidates. The phrases chosen as cue phrases are selected principally using a criterion of predictivity.
Data Set & Cross Validated Score & Single Best Score \\ 
50k, unclustered & 56.35% & 60.67% \\ 
50k, clustered 42 tags & 61.29% & 65.80% \\ 
as above, plus utt. length models & 65.71% & 68.78% \\ 
as above, plus <start>,<end> tags & 66.41% & 69.53% \\ 
as above, plus interrupted utterances & 68.42% & 71.98% \\ 

Figure 4: Experiments with 50k data set

tivity, which is the extent to which the presence of a certain n-gram in an utterance is predictive of it having a certain dialogue act category. For an n-gram n and dialogue act d, this corresponds to the conditional probability: \( P(d|n) \), a value which can be straightforwardly computed. Specifically, we compute all n-grams in the training data of length 1–4, counting their occurrences in the utterances of each DA category and in total, from which the above conditional probability for each n-gram and dialogue act can be computed. For each n-gram, we are interested in its maximal predictivity, i.e. the highest predictivity value found for it with any DA category. This set of n-grams is then reduced by applying thresholds of predictivity and occurrence, i.e. eliminating any n-gram whose maximal predictivity is below some minimum requirement, or whose maximal number of occurrences with any category falls below a threshold value. The n-grams that remain are used as cue phrases. The threshold values that were used in our experiments were arrived at empirically, by conducting a series of experiments at varying levels of threshold and frequency.

4.3 Using Cue Phrases in Classification

The selected cue phrases are used directly in classifying further utterances in the following manner. To classify an utterance, we identify all the cue phrases it contains, and determine which has the highest predictivity of some dialogue act category, and then that category is assigned. If multiple cue phrases share the same maximal predictivity, but predict different categories, one category is assigned arbitrarily. If no cue phrases are present, then a default tag is assigned, corresponding to the most frequent tag within the training corpus.

4.4 Experimental cases

For each of the three data sets, we performed five different experiments, whose results are reported in Figures 3–5. The five different experimental cases are described following.

Case 1: baseline

For these experiments, the classification approach just described was applied using the full 220 element tag set from the SWITCHBOARD corpus. This provides us with baseline figures of tagging accuracy. Applied to the 4k data set, the approach yields an average tagging accuracy of 51.83%, which compares against a baseline accuracy of 36.5% from applying the most frequently occurring tag in the SWITCHBOARD data set (which is sd — statement). Applied to the medium data set, the approach yields an average tagging accuracy of 54.5%, which compares to 33.4% from using the most frequent tag. Finally, applied to the large data set, we produced an average tagging accuracy of 55.82%, compared to a baseline of 36%.

Case 2: clustered tag set

For these experiments, we used the clustering of labels proposed by Jurafsky et al. (1998), which maps the full 220 DA labels in the 42 larger classes shown in Figure 2. This move produced a significant improvement in performance, around 5% in all cases. For the 4k data set, average tagging accuracy rose to 56.47% (an improvement of 4.64%). For the 50k data set, the score was 61.29% (an improvement of 4.94%), and for the 202k data set we achieved 60.73% (4.91%).
Case 3: utterance length models
For this case, we trained models sensitive to utterance length. In particular, we grouped training utterances into those of length 1, those with lengths 2–4, and those of length 5+, and produced separate modules for each group. We hoped that this move would provide better classification for dialogue acts whose realisation was skewed over, for instance, short utterances like ‘okay’. On the whole, the introduction of such models lead to an increase in tagging accuracy of around 4%, except in the case of the 4k set, where data sparsity was more of an issue.

Case 4: position specific cues
Further experiments suggest that we can improve this score. We introduced start and finish tags to each utterance - to capture position specific information for particular cues. For example ‘start okay’ effectively identifies occurrence of word ‘okay’ as first word in the utterance. The effects of these additions can be seen in the tables, but in summary, the position specific cues added a further percentage point.

Case 5: interrupted utterances
In addition to the dialogue act mark-up of the corpus, there were several annotations relating directly to the dysfluencies encountered in the data. These are outlined in (Meteer, 1995) and are reported in (Stolcke et al., 2000) as not being considered in their experiments as they fall outside the 42 tag set. The most important of these is the dialogue act ‘+’, which indicates an utterance which was interrupted by the other speaker, an example of which can be seen in Figure 6.

We saw that a lot of potentially useful word data was being ignored. The ‘+’ tag occurs around 16,000 times in the 202k corpus, around 8% of total annotations. One approach to utilise this data would be to ‘reconnect’ the segmented data, i.e. appending any data assigned tag ‘+’ to the last utterance by the same speaker. Clearly this approach has its limits — and such a corpus would lose important sequence information (such as the effect of back-channels on the conversation). However, as a pre-processing step, it is worth exploring. Doing so gave us both our highest cross-validated score of 69.09% and our highest single score of 71.98%.

5 N-Best Dialogue Act Classification
All experiments up to the present have tried to select the single best-fit candidate dialogue act tag for an utterance. As this could be seen as a first step before possible refinement by some higher level process, we wanted to investigate the possibility of selecting some list, possibly ranked, of potential dialogue acts. The higher level process, perhaps itself a machine learning algorithm or dialogue manager, could choose among some limited selection of possible acts based on additional information outside the utterance itself, such as dialogue context.

This would address the problem of being unable to resolve some ambiguity on the basis of surface realisation. For example, the utterance ‘okay’ can be either a back-channel or an accept/confirm, it depends entirely on the context. If we can represent such an ambiguity to a higher level process, a restricted choice can be made based on contributory factors.

Our most recent experiment shows interesting promise. We built a classifier using the 45k utterance training set, and tested it on the 5k
utterance test set. However, rather than attempting to find the single best match from the classifier, we tagged each utterance with the top 5 possible dialogue acts, as indicated by the classifier on the basis of the predictivity of the n-grams the utterance contained. On a cross-validation of the corpus, we calculated that 86.74% of the time the correct dialogue act was contained in the 5-best output of the classifier. In order to create some baseline measure, this experiment was repeated using the top 5 n-grams occurring by frequency in the SWITCHBOARD corpus. The number of times the correct dialogue act occurred in the top 5 was 71.09%.

The upper score here would define some theoretically attainable upper limit of performance, but one where the complexity of choice is reduced from 42 to 5.

6 Discussion

Combining all features for simple dialogue act tagging, we obtain a cross validated score of 69.09% over the larger, 202k data set. Our highest single run score was 71.98%, using the 50k data set. It is difficult to compare our results directly with those of (Stolcke et al., 2000), given that did not use a cross-validation approach, but even so it is striking that our cross-validated score comes so close to their result given their use of a much more complex language modelling approach, that exploits also inter-utterance information. It is furthermore possible that their choice of test data was a lucky one, i.e. one giving higher scores than would arise with results averaged in cross-validation.

We have shown that a simple dialogue act tagger can be created that uses just intra-utterance cues for classification. This approach performs surprisingly well given its simplicity. One of the prime motivators for using this approach was to remove a large number of word n-grams from the feature set of machine learning algorithms. By doing so we are hopeful that we can use a wider range of machine learning approaches for this task than has presently been tried. Finally, by analysing the n-best approach to tagging, we have demonstrated that a naive classifier can present a list of ranked possible alternatives, which could be used by some later, higher level structure, such as a dialogue manager, to make informed choices in the evaluation of utterances.

7 Future Work

Clearly one next step is to pass these results to some machine learning algorithm, to exploit inter-utterance relationships. In the first instance, Transformation-Based Learning (TBL) will be investigated, but the attractiveness of this approach to previous researchers (Samuel, Carberry and Vijay-Shanker, 1998; Lager and Zinovjeva, 1999) was in part the tolerance of TBL to a potentially large number of features. we will use our naive classification method to pass as a single feature our best-first guess. Additionally, we will look to include the n-best list of DAs as a feature to a machine learning algorithm. If we can restrict the mechanism to assigning only those DAs which occur in our n-best list, we hope to improve tagging performance.

We can further improve our naive classification method. We are concerned that the model we generate displays over-fitting with respect to the training data, so to counter this, we intend to split training data into two parts - training and validation. after training is complete, we will validate on the second part of the data, to prune the model of those elements caused
by over-fitting. Then the model will be tested. Hopefully this will lead to better performing, more general models.

As a test, we will also start to tag the VERB-MOBIL corpus data. An interesting area of investigation is to what extent models trained on one set of data can be used to tag data from a different domain and conversational style.

Finally, we aim to apply these techniques to a new corpus collected for the AMITIES project, consisting of human-human conversations recorded in the call centre domain. We hope that the techniques outlined here will prove a useful first step in creating automatic service counters for call centre applications.

8 Acknowledgements

This paper is based on work supported in part by the European Commission under the 5th Framework IST/HLT Program, and by the U.S. Defense Advanced Research Projects Agency.

References


