Error Analysis of Dialogue Act Classification

Nick Webb, Mark Hepple, and Yorick Wilks

Natural Language Processing Group
Department of Computer Science
University of Sheffield, UK
{n.webb,m.hepple,y.wilks}@dcs.shef.ac.uk

Abstract. We are interested 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. We have built a simple dialogue act classifier based on purely *intra-utterance* features — principally word n-gram cue phrases. Although such a classifier performs surprisingly well, rivalling scores obtained using far more sophisticated language modelling techniques for the corpus we address, we want to understand further the issues raised by this approach. We have performed an error analysis of the output of our classifier, with a view to casting light both on the system's performance, and on the DA classification scheme itself.

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. In common with the work of Samuels *et al.* [1], we have automatically detected word n-grams in a corpus that might serve as useful cue phrases, potential indicators of dialogue acts. 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, that is, makes no use of the relationship between utterances, such as preceding DA labels.

Having built such a classifier [2], we were surprised that the results we obtain rival the best results achieved on our target corpus, in work by Stolcke *et al.* [3], 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. In order that we might better understand the performance of our classifier, we decided to perform a simple error analysis - looking at which of the categories in our corpus are most consistently tagged correctly, and which are not. In Section 2 of this paper, we present previous work in DA modelling. A brief overview of our previous classification experiments is presented in Section 3. In Section 4 we outline the error analysis performed using this classifier, and in the light of this, an experiment in collapsing some of the target categories for the corpus is reported in Section 5. We end with some discussion and an outline of intended further work.

2 Related Work

One approach utilised for dialogue act tagging is that of n-gram language modelling, exploiting principally ideas drawn from the area of speech recognition. For example, Reithinger and Klesen [4] 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.* [3] apply a somewhat more complicated HMM method to the SWITCHBOARD corpus, one which 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 *et al.* [5], 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. One 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'.

Dialogue Act	*% of corpus*	*% accuracy*	*Dialogue Act*	*% of corpus*	*% accuracy*
statement-non-opinion	36%	92%	action-directive	0.4%	25%
acknowledge	19%	97%	collaborative completion	0.4%	0%
statement-opinion	13%	26%	repeat-phrase	0.3%	0%
agree-accept	5%	31%	open-question	0.3%	67%
abandoned	5%	55%	rhetorical-questions	0.2%	0%
appreciation	2%	86%	hold before answer	0.2%	33%
yes-no-question	2%	51%	reject	0.2%	0%
non-verbal	2%	100%	negative non-no answers	0.1%	14%
yes answers	1%	0%	signal-non-understanding	0.1%	0%
conventional-closing	1%	47%	other answers	0.1%	0%
uninterpretable	1%	55%	conventional-opening	0.1%	50%
wh-question	1%	46%	or-clause	0.1%	0%
no answers	1%	5%	dispreferred answers	0.1%	0%
response acknowledgement	1%	0%	3rd-party-talk	0.1%	0%
hedge	1%	64%	offers, options commits	0.1%	0%
declarative yes-no-question	1%	3%	self-talk	0.1%	0%
other	1%	0%	downplayer	0.1%	0%
backchannel in question form	1%	19%	maybeaccept-par	< 0.1%	0%
quotation	0.5%	0%	tag-question	< 0.1%	0%
summarisereformulate	0.5%	0%	declarative wh-question	< 0.1%	0%
affirmative non-yes answers	0.4%	0%	apology	< 0.1%	50%

Fig. 1. SWITCHBOARD dialogue acts, by occurrence and tagging accuracy

3 Simple DA Classification

In Webb *et al.* [2], we describe our simple approach to DA classification, based solely on intra-utterance features, together with our evaluation experiments. A key aspect of our approach is the selection of word n-grams to use as cue phrases in tagging. Samuel *et al.* [1] 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.

3.1 Experimental corpus

For our experiments, we used the SWITCHBOARD data set of 1,155 annotated conversations, which together comprise in the region 205k utterances. The dialogue act types for this set can be seen in Jurafsky *et al.* [6]. The corpus is annotated using an elaboration of the DAMSL tag set [7], involving 50 major classes, together with a number of diacritic marks, which combine to generate 220 distinct labels. Jurafsky *et al.* [6] propose a clustering of the 220 tags into 42 larger classes and it is this clustered set used both in the experiments of Stolcke *et al.* [3], and those reported here. The 42 DA classes can be seen in Figure 1. We used 198k utterances for training and 4k for testing, with pre-processing to remove all punctuation and case information, in common with Stolcke *et al.* [3] in order that we might compare figures. Some of the corpus mark-up, such as filler information described in the paper by Meteer [8], was also removed.

Our experiments use a cross-validation approach, with results being averaged over 10 runs. For our data, the test set is much less than a tenth of the overall data, so a standard ten-fold approach does not apply. Instead, we randomly select dialogues out of the overall data to create ten subsets of around 4k utterances for use as test sets. 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.* [3].

3.2 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*, 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: $P(d \mid n)$, 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.

3.3 Using Cue Phrases in Classification

The selected cue phrases are used directly in classifying previously unseen 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, the category indicated by the phrase with the *highest* frequency count is assigned. If no cue phrases are present, then a default tag is assigned, corresponding to the most frequent tag within the training corpus.

3.4 Experimental cases

In our previous work [2] we performed five different experiments using a variety of simple methods for pre-processing the data. Our best reported figures on the 202k utterance corpus are a cross-validated score of 69.09%, with a single high score of 71.29%, which compares well with the (non-cross-validated) 71% reported in Stolcke *et al.* [3].

In each experiment, there are two important variables used to select n-grams as potential cue phrases - the frequency of occurrence of each n-gram, and how predictive an n-gram is of some dialogue act. In more recent work [9], we have shown that we can select these scores automatically, using a validation set separate from our test data, without suffering a significant decrease in tagging performance. During these experiments, we observed some dialogue act categories that seemed to be most easily confused - where utterances of one category are consistently incorrectly tagged as being of a second category. The focus of this paper is to perform an explicit error analysis to determine empirically which categories are most confused by our classification method. By further inspection of the errors, perhaps there are issues within our model which can be adjusted to attain better performance, or perhaps we have reached the limit possible given only intra-utterance classification models, and the resulting ambiguities can be solved only with reference to dialogue context.

4 Error Analysis

Taking our best performing DA model from Webb *et al.* [2], we performed an error analysis, calculating the accuracy of recognition for each dialogue act in the corpus. The results of a single, non-cross-validated run, using a test set of 3279 utterances and a training set of some 198k utterances, can be seen in Figure

Dialogue Act Name	*count*	*% accuracy*
statement-opinion	469	26%
Incorrectly tagged as:	*count*	*% error*
appreciation	4	0.9%
abandoned	13	2.8%
yes-no-question	1	0.2%
hedge	1	0.2%
acknowledge	2	0.4%
conventional-closing	1	0.2%
statement-non-opinion	315	67.2%
acknowledge-accept	11	2.3%
wh-question	1	0.2%

Fig. 2. Single category error analysis

1. There are some interesting points to note here. We can see that 'statements-non-opinion' score very highly 92% recognition accuracy), but that 'statements-opinion' score far lower (26%). What could be worrying, for automatic dialogue systems, is the low recognition rate for important categories such as 'yes-answers' (0%) and 'no-answers' (5%). Some of these low scores can be attributed to sparse amounts of training data for specific dialogue acts - given their low frequency of occurrence in the SWITCHBOARD corpus. Others can be easily explained in terms of the lexicalisations which realise the utterances. We will look at this in more detail later.

Alone, these figures do not necessarily help us identify areas of the classification mechanism on which we should focus for improvement. For each occurrence of a dialogue act which we tagged incorrectly, we noted which tag was used in error. For example, the tag 'statement-opinion' occurs 469 times in the chosen test data, of which we correctly tagged 120, or 26%. Of the incorrect tags assigned to 'statement-opinion' utterances, the scores for tagging with other DAs can be seen in Figure 2, along with the proportional scores calculated by dividing the number of times an incorrect tag is used for a specific category, by the number of times the correct category occurs in the corpus. It seems clear that the significant score here is the number of times that a 'statement-opinion' utterance is tagged as a 'statement-non-opinion'. We determined that this proportional score is one useful discriminator for selecting interesting, regularly confused DA tags. We chose to look at only those tags where 50% or more of the proportion of errors to total occurrences are tagged as a single incorrect category.

An equally important factor is the number of occurrences of the DA tag in question. It makes sense in the first instance to concentrate on those DAs whose count was significant - i.e. those where correcting errors in classification would have a statistically significant effect on classifier performance. We chose to concentrate on those DAs whose occurrence in the test data is higher than 40 - the equivalent to an effective 1% gain in classifier performance, if all instances are tagged correctly. Interestingly, there are only two instances where both criteria of significant count and significant proportional errors are fulfilled. The first of these is as already mentioned, the case of 'statement-opinion' being incorrectly tagged as 'statement-non-opinion'. The second is the case of 'agree-accept' being tagged as 'acknowledge' (there are 228 instances of 'agree-accept' in the test data, of which 70 were tagged correctly; of the 158 errors, 140 were tagged as 'acknowledge', 61.4% of all instances). We shall examine both cases in turn.

For the confusion regarding 'statement-opinion', we first look at the tagging guidelines for the SWITCHBOARD corpus, laid out in Jurafsky et al. [6]. They themselves are unable to ascertain if the distinction between the categories is fruitful. Having trained separate tri-gram models on the two sets, Stolcke *et al.* claim these tri-gram models look somewhat distinct, and yet found that this distinction was very hard to make by human labellers. Jurafsky *et al.* report that this distinction accounted for a large proportion of their inter-labeller disagreements. They provide 'clue' phrases, which may be present in a 'statement-opinion' utterance. These include: *'I think'; 'I believe; 'It seems'* and *'It's my opinion that'.* Looking at the n-grams created from the entire corpus, we can start to identify some potential problems. 'I think' is a common n-gram, occurring more than 6250 times. However, while some 63% of those occur in utterances tagged with 'statement-opinion', there is still a large margin of error. 31% of the remaining n-grams occur in 'statement-non-opinion' utterances. This position is much the same with 'it seems' (472 total instances, 307 as 'statement-opinion' (65%), 144 as 'statement-non-opinion' (31%). In these cases, although the 'clue phrases' are clearly indicative of 'statement-opinion', if some other, more highly predictive n-gram is present, it's possible that the presence of such clues will be ignored. It is even worse with respect to 'I believe', which occurs 190 times in total, but where 88 (46%) of those predict 'statement-opinion', 91 (48%) occur in 'statement-non-opinion' utterances. The only one of Jurafsky's examples to fare well is 'it's my opinion that', but as this occurs only once in the entire corpus, is of limited use. This investigation bears out the argument that labellers had extreme difficultly in differentiating between these two categories. There is then a substantial argument here that if this is a hard category for human labellers to separate, perhaps there should not be two, distinct categories.

The second problem category, where 'agree-accept' can often be tagged as 'acknowledge', is more straightforward to understand. By looking at a sample of the utterances coded in each category, we can see that, as might be expected, they have substantially similar lexicalisations. Both are represented often by 'yeah', 'yes' and 'right'. According to the labellers manual [6], there are several contextual issues which may help to disambiguate the two. This raises an important point. Since this far we have concerned ourselves only with intra-utterance features, we are unable to disambiguate some of the categories at this stage. We hope that higher level processes, perhaps powered by machine learning alorithms, may enable to us to leverage the context of surrounding utterances in our classification. We speculate that a machine learning approach, using context, might do better at disambiguating between ‘agreement-accept’ and ‘acknowledge’, but not do significantly better for ‘statement-opinion’ and ‘statement-non-opinion’.

5 Merging categories

As we have shown, the categories ‘statement-opinion’ and ‘statement-non-opinion’ are often confused. This split of the STATEMENT category is one that Jurafsky et al. created for the SWITCHBOARD corpus, as there is no such distinction made in the DAMSL [7] coding schema from which the annotation set for the SWITCHBOARD corpus is derived. In order to test the performance of a system where such a distinction was not made, we created a version of the corpus where all instances of both ‘statement-opinion’ and ‘statement-non-opinion’ were *replaced* by the single category ‘statement’. The results from the error-analysis would seem to indicate that there should be an almost 10% improvement in our classification accuracy. In previous work, we reported a best cross-validated score of 69.09% (with a high score of 71.29%) [2]. After repeating the cross-validation experiment on the new corpus, we achieve a score of 76.73%, with a high of 78.58%, in both cases a gain of over 7%.

`Speaker A: DA="statement-non-opinion":` **but I also believe that the earth is a kind of a self-regulating system**

Fig. 3. An example utterance incorrectly labelled

Another possible solution to this problem is to use the phrases suggested by Jurafsky *et al.*, and their variants, to create a distinct set of utterances, all of which *should* be labelled at ‘statement-opinion’. This would correct the error indicated in Figure 3. Alternatively, when we merge the ‘statement-opinion’ and ‘statement-non-opinion’ utterances into a single category, we propose a separate indicator of whether an utterance contains a lexical indicator of opinion. This would make annotation easier, in that when clear evidence of opinion was identified, this information could be added to the basic ‘statement’ annotation.

6 Discussion, Future Work

The task of labelling spoken, conversational data is clearly complex. Our error analysis has shown that some categories are difficult for humans and machines to separate. Perhaps this can be turned into a mechanism whereby we can have some automatic measure of the efficiency of coding schemes. One possible limitation of our error analysis is the question of whether the problems faced are specific to our classification approach. If the problems we report are common across a range of tagging approaches, this presents a stronger argument for merging categories. Stolcke *et al.* and Jurafsky *et al.* both indicate difficulties with the categories we identify as problematic.

We have shown that a simple dialogue act tagger can be created that uses solely intra-utterance cues for classification. This approach performs surprisingly well given its simplicity. However, in order to improve the performance of our classifier still further, it is clear that we need to make use of features outside of the individual utterance - such as DA sequence information. 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 [5] was due in part to the tolerance of TBL to a potentially large number of features. We will use our classification method to pass as a single feature our suggested category for each utterance, without the need to represent the large set of word n-grams in the learning algorithm's feature set. If this proves successful we can use a far larger set of possible machine learning approaches to advance our classification performance.

References

1. Samuel, K., Carberry, S., Vijay-Shanker, K.: Automatically selecting useful phrases for dialogue act tagging. In: Proceedings of the Fourth Conference of the Pacific Association for Computational Linguistics, Waterloo, Ontario, Canada. (1999)
2. Webb, N., Hepple, M., Wilks, Y.: Dialogue Act Classification Based on Intra-Utterance Features. In: Proceedings of the AAAI Workshop on Spoken Language Understanding. (2005)
3. Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C.V., Meteer, M.: Dialogue act modeling for automatic tagging and recognition of conversational speech. In: Computational Linguistics 26(3), 339–373. (2000)
4. Reithinger, N., Klesen, M.: Dialogue act classification using language models. In: Proceedings of EuroSpeech-97. (1997)
5. Samuel, K., Carberry, S., Vijay-Shanker, K.: Dialogue act tagging with transformation-based learning. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal (1998)
6. Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., Ess-Dykema, C.V.: Switchboad discourse language modeling project final report. Research Note 30, Center for Language and Speech Processing, Johns Hopkins University, Baltimore (1998)
7. Core, M.G., Allen, J.: Coding dialogs with the DAMSL annotation scheme. In: AAAI Fall Symposium on Communicative Action in Humans and Machines, MIT, Cambridge, MA (1997)
8. Meteer, M.: Dysfluency annotation stylebook for the switchboard corpus. Working paper, Linguistic Data Consortium (1995)
9. Webb, N., Hepple, M., Wilks, Y.: Empirical determination of thresholds for optimal dialogue act classification. In: Proceedings of the Ninth Workshop on the Semantics and Pragmatics of Dialogue. (2005)