Empirical determination of thresholds for optimal 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 present recent experiments which build on our 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. We describe a simple dialogue act classifier based on purely *intra-utterance* features — principally 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. The approach requires the use of thresholds effecting the selection of n-gram cues, which have previously been manually supplied. We here describe a method of automatically determining these thresholds to optimise classifier performance.

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 et al., 1999), we wanted to detect automatically word n-grams in a corpus that might serve as potentially 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. The core of this paper investigates whether the crucial values for predicitivity of cue phrases can be determined empirically using a validation set of data, held out from evaluation. In a recent paper (Webb et al., 2005), we report early results of experiments evaluating our simple approach to

classification on the SWITCHBOARD corpus, using manually pre-set thresholds for our key variables. Surprisingly, the results we obtain rival the best results achieved on that corpus, in work by Stolcke *et al.* (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.

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 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., 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 Simple DA Classification

In previous work, we describe our simple approach to DA classification, based on intra-utterance features, together with our experiments to evaluate it (Webb et al., 2005). A key aspect of the approach is the selection of word n-grams to use as cue phrases in tagging. (Samuel et al., 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.

Dialogue Act	*% of corpus*	*Dialogue Act*	*% of corpus*
statement-non-opinion	36%	action-directive	0.4%
acknowledge	19%	collaborative completion	0.4%
statement-opinion	13%	repeat-phrase	0.3%
agreeaccept	5%	open-question	0.3%
abandoned	5%	rhetorical-questions	0.2%
appreciation	2%	hold before answer	0.2%
yes-no-question	2%	reject	0.2%
non-verbal	2%	negative non-no answers	0.1%
yes answers	1%	signal-non-understanding	0.1%
conventional-closing	1%	other answers	0.1%
uninterpretable	1%	conventional-opening	0.1%
wh-question	1%	or-clause	0.1%
no answers	1%	dispreferred answers	0.1%
response acknowledgement	1%	3rd-party-talk	0.1%
hedge	1%	offers, options commits	0.1%
declarative yes-no-question	1%	self-talk	0.1%
other	1%	downplayer	0.1%
backchannel in question form	1%	maybeaccept-par	$< 0.1\%$
quotation	0.5%	tag-question	$< 0.1\%$
summarisereformulate	0.5%	declarative wh-question	$< 0.1\%$
affirmative non-yes answers	0.4%	apology	$< 0.1\%$

Figure 1: SWITCHBOARD dialogue acts

3.1 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 et al., 1997). Altogether these 1,155 conversations comprise in the region of 205k utterances.

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 1, and it is this clustered set used both in the experiments of (Stolcke et al., 2000), and those reported here.

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., 2000) in order that we might compare figures.

Some of the corpus mark-up, such as filler information described in (Meteer, 1995), 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., 2000).

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. The n-grams that remain are used as cue phrases. It should be obvious that the levels of these two thresholds, frequency and predictivity, are crucial to the performance of the system.

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, 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.

3.4 Experimental cases

In previous work (Webb et al., 2005) we performed five different experiments using a variety of simple word processing techniques. The model which gained the best results used a corpus clustered into 42 dialogue act classes, had special tags marking the beginning and end of each utterance, had models trained for different lengths of user utterances and removed some of the effects of disfluencies from the corpus. 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., 2000).

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 the notion of how predictive a particular n-gram is of some dialogue act.

The values of these variables were set in an arbitrary manner, selecting most likely candidates through prior knowledge of experiments. In the experiments reported in (Webb et al., 2005), these are a minimum frequency count of 2, and minimum predictivity score of 30%. N-gram cues with scores lower than these thresholds were discarded from the possible set used for classification.

This approach has a number of inherent problems. First, we do not know if there are some other values which will work better. The scores we used were chosen following extensive work with a 50k utterance training set - it is possible these pre-set threshold values would no longer be optimal when used with larger training sets.

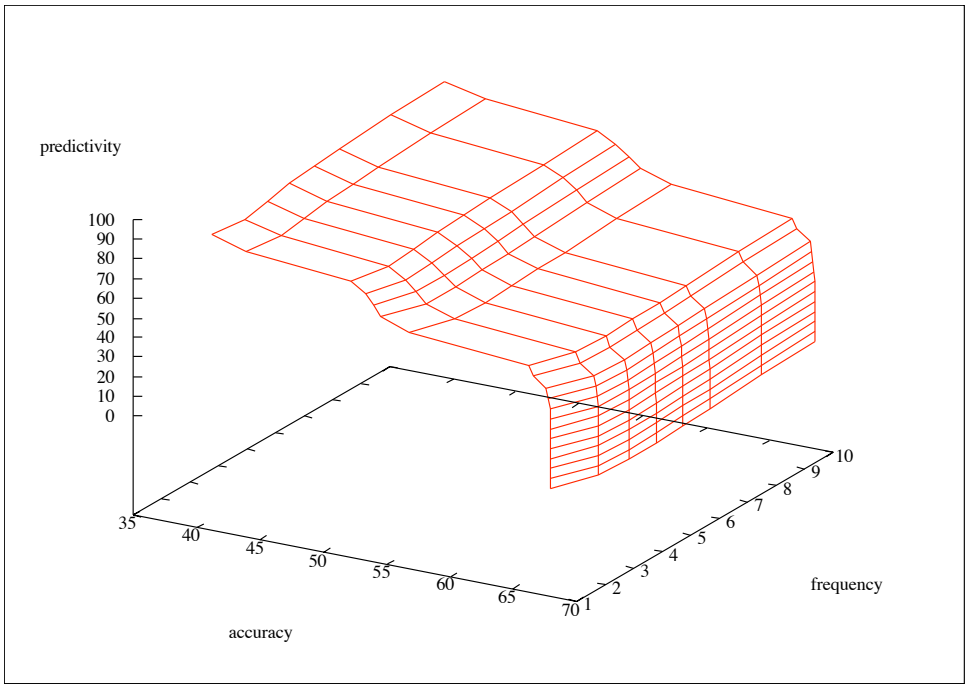

Figure 2: Effects of predictivity and frequency on tagging accuracy

Secondly, these values were chosen for their ability to perform well over the test data. Such an approach undermines our attempts to establish a baseline for classification performance. Our subseqeuent experiments aim to address these problems directly.

3.5 Exhaustive Thresholds

To address the two concerns we sought to develop a method that would determine thresholds automatically, as part of the training process, through the use of a validation set. As a prelude to this, we investigated how the performance of the classification approach interacts with the selection of thresholds, by computing performance results at an exhaustive range of threshold values.

For this search we computed scores for frequency count thresholds of 1, 2, 3, 4, 5, 6, 8 and 10. For predictivity, all scores from 0 to 100% were used in steps of 5%, for each of the possible frequency cut-offs.

Figure 2 shows the effect on tagging accuracy of varying thresholds. A quick interpretation of this graph shows that the classifier performs well with minimum predictivity thresholds of 40% and below, but falls rapidly for thresholds above that value. Although the classifier performs optimally with a frequency threshold around 2 or 3, the behaviour is tolerant of higher thresholds.

As can be seen in Figure 3, the best cross validated accuracy scores occur at a frequency count of 3, minimum predictivity of 35%. This score is higher than our manually selected thresholds of frequency 2, predictivity 30%, although the effective gain is 0.17%

Additionally, this single highest score oc-

Freq	*Pred*	*Cross Validated Score*	*Single Best Score*
2	30	69.48%	74.89%
3	30	69.65%	**74.95%**
3	35	**69.65%**	74.92%

Figure 3: Experiments with 202k data set

curs at 30% predictivity, although again the difference is extremely low, at 0.06%. It is worth noting that the figures quoted here for both cross-validation and single highest score are greater than our best published figures to date, and the highest score is 3.95% higher than that reported in (Stolcke et al., 2000).

3.6 Validation Model

We recognise that selecting thresholds manually by performance on the test set may not be a robust method for this task. To counter this, we split training data into two parts - training and validation. After training is complete, we will validate on the second part of the data, to automatically select the best values for minimum frequency and predictivity counts. This directly addresses the original problem of setting values based on the test data.

Experimentally, we now take the 198k utterance training set, and take 10% (around 20k utterances) to use for validation, a set distinct from the 4k utterances used for testing. We derive n-grams from the 178k training set, then do exhaustive testing over the validation set, using the range of variables described in the previous experiment. These experiments select the best performing combination of frequency and predictivity scores which are then used when applying the n-grams to the test set. We repeat this 10 times, using a random selection of dialogues for both the validation and testing data sets. In each case, we also tag the test data using our original, arbitrary values of frequency 2, predictivity 30%, to establish some kind of baseline.

The average frequency count selected by our automatic method is 2.9, average minimum predicitivity of 32.5%. The cross-validated tagging accuracy when classifying using these automatically selected thresholds is 67.44% (with a high score of 70.31%). This compares favourably to the cross-validated score of 67.49% (high score 70.72%) obtained using our static, manually prescribed thresholds on the same data splits. These results are perhaps not surprising given the previous experiment, which seems to demonstrate a broad range of values for these thresholds over which tagging accuracy is largely unaffected.

These overall cross-validated scores seem to be down on other reported scores - this could be due in part to the loss of training data caused by the creation of the validation set. However it is encouraging to see that we can use the validation data to select scores which perform well over the test data.

4 Discussion, Future Work

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. The model appears to be robust, given that there is a range of possible values which combine to give good tagging accu-

racy scores. We are able to determine the settings for these variables independently from the test data.

Future work include a thorough investigation of the effects of the amount of data available for training, and the most effective size of validation set. Further, an error analysis of the data, to determine which dialogue act classes are most easily confused, would be interesting.

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 et al., 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.

References

Mark G. Core and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, MIT, Cambridge, MA.

Barbara Grosz and Candace Sidner. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 19(3).

Julia Hirschberg and Diane Litman. 1993. Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics*, 19(3):501–530.

D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. Van Ess-Dykema. 1997. Automatic detection of discourse structure for speech recognition and understanding. In *Proceedings of the 1997 IEEE Workshop on Speech Recognition and Understanding, Santa Barbara.*

Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1998. Switchboad discourse language modeling project final report. Research Note 30, Center for Language and Speech Processing, Johns Hopkins University, Baltimore.

Torbjörn Lager and Natalia Zinovjeva. 1999. Training a dialogue act tagger with the μ-TBL system. In *Proceedings of the Third Swedish Symposium on Mulitmodal Communication, Linköping University Natural Language Processing Laboratory.*

Marie Meteer. 1995. Dysfluency annotation stylebook for the switchboard corpus. Working paper, Linguistic Data Consortium.

Norbert Reithinger and Martin Klesen. 1997. Dialogue act classification using language models. In *Proceedings of EuroSpeech-97.*

Ken Samuel, Sandra Carberry, and K. Vijay-Shanker. 1998. 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.

Ken Samuel, Sandra Carberry, and K. Vijay-Shanker. 1999. Automatically selecting useful phrases for dialogue act tagging. In *Proceedings of the Fourth Conference of the Pacific Association for Computational Linguistics, Waterloo, Ontario, Canada.*

A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. In *Computational Linguistics 26(3), 339–373.*

Nick Webb, Mark Hepple, and Yorick Wilks. 2005. Dialogue Act Classification Based on Intra-Utterance Features. Research Memorandum CS-05-01, Department of Computer Science, University of Sheffield.