Automatic Extraction of Cue Phrases for Cross-Corpus Dialogue Act Classification

Nick Webb and **Michael Ferguson**
ILS Institute, SUNY Albany
`nwebb@albany.edu, ferguson@cs.albany.edu`

Abstract

In this paper, we present an investigation into the use of cue phrases as a basis for dialogue act classification. We define what we mean by cue phrases, and describe how we extract them from a manually labelled corpus of dialogue. We describe one method of evaluating the usefulness of such cue phrases, by applying them directly as a classifier to unseen utterances. Once we have extracted cue phrases from one corpus, we determine if these phrases are general in nature, by applying them directly as a classification mechanism to a different corpus to that from which they were extracted. Finally, we experiment with increasingly restrictive methods for selecting cue phrases, and demonstrate that there are a small number of core cue phrases that are useful for dialogue act classification.

1 Motivation

In this paper we present a recent investigation into the role of linguistic cues in dialogue act (DA) classification. Dialogue acts (Bunt, 1994) are annotations over segments of dialogue that characterise the function of those segments. Linguistic cues, which can take many forms including lexical and syntactic structures, are features that can serve as useful indicators of discourse structure (Hirschberg and Litman, 1993; Grosz and Sidner, 1986). In prior work, several researchers have shown that cue phrases can be a powerful feature for DA classification (Samuel et al., 1999; Webb et al., 2005a). Webb and Liu (2008) have previously shown that cue phrases automatically extracted from one corpus can be used to classify utterances from a new corpus. We take this

approach and apply it to two established corpora with manually encoded dialogue act annotations, to investigate both the existence and the usefulness of cue phrases shared between the two corpora.

2 Related Work

In parallel with the increased availability of manually annotated dialogue corpora there has been a proliferation of literature detailing dialogue act labelling as a classification task. Prior work describes the selection of features from the corpus (including word n-grams, cue phrases, syntactic structures, dialogue history and prosodic cues) which are then passed to some machine learning algorithm. Most studies have concentrated on a single corpus, and optimised feature selection and learning algorithm accordingly. In this work we focus on two corpora, Switchboard and ICSI-MRDA, and discuss prior classification efforts relating to these two corpora.

2.1 Switchboard Corpus

The Switchboard corpus contains a large number of approximately 5-minute conversations between two people who are unknown to each other, who were asked to converse about a range of everyday topics with little or no constraint. The DA annotated portion of the Switchboard corpus (Jurafsky et al., 1997) consists of 1155 annotated conversations, containing some 225,000 utterances, of which we use 200,000 utterances, the rest being held out for separate experiments. The dialogues are annotated with a non-hierarchical variant of the DAMSL annotation scheme (Core et al., 1999). The resulting Switchboard-DAMSL annotation was a set of more than 220 distinct labels. To obtain enough data per class for statistical modelling purposes, a clustered tag set was devised, which distinguishes 42 mutually exclu-

sive DA types. Classification over the Switchboard corpus has been demonstrated using Decision Trees (Verbree et al., 2006), Memory-Based Learning (Rotaru, 2002) and Hidden Markov Models (HMM) (Stolcke et al., 2000). The work of Stolcke et al. (2000) is often cited as the best performing, achieving a classification accuracy of 71% over the 42 labels, although there is no cross-validation of these results. The approach of Stolcke et al. (2000) combines HMM modelling of utterances with a tri-gram model of DA sequences. Webb et al. (2005a) report a slightly lower cross-validated score (of 69%) containing an individual classification high of 72%, using an intra-utterance, cue-based classification model.

2.2 ICSI-MRDA Corpus

Like the Switchboard corpus, the ICSI Meeting Room DA (MRDA) corpus (Shriberg et al., 2004) was annotated using a variant of the DAMSL tagset, similar but not identical to the Switchboard-DAMSL annotation. The differences (and a translation between the two sets) can be seen in Shriberg et al. (2004). The underlying domain of the dialogues in the ICSI-MRDA corpus was that of multi-party meetings, with multiple participants discussing an agenda of items in a structured meeting. This application required the introduction of new tags specifically for this scenario, such as a label introduced to indicate when an utterance was used to take control of the meeting. The ICSI-MRDA corpus comprises 75 naturally occurring meetings, each around an hour in length. The section of the corpus we use consists of around 105,000 utterances. For each utterance in the corpus, one general tag was assigned, with zero or more additional specific tags. Excluding non-labelled cases, there are 11 general tags and 39 specific tags resulting in 1,260 unique dialogue acts used in the annotation. As with the Switchboard corpus, processing steps were introduced that compressed the number of unique DAs to 55. In later work, the dimensionality was further reduced, resulting in a subset of just 5 labels.

Over the ICSI-MRDA corpus, we also see DA classification efforts using Decision Trees (Verbree et al., 2006) and Memory-Based Learning (Lendvai and Geertzen, 2007), in addition to

Graph Models (Ji and Bilmes, 2006) and Maximum Entropy (Ang et al., 2005). Comparatively few approaches have been applied to the 55-label annotated corpus, with most choosing to focus on the 5-label clustering, presumably for the resulting increase in score. When Ji and Bilmes (2005) apply a Graph Model to the 55 category corpus, they achieve a classification accuracy of 66%. However, when they apply the exact same method to the 5-label corpus (Ji and Bilmes, 2006), classification accuracy is boosted to 81%. The best reported classification score on the the 5-label version of the corpus is reported by Verbree et al. (2006), who achieve 89% classification accuracy by modelling the words of the utterance, the DA history and some orthographic information (such as the presence of question marks).

It remains very difficult to directly compare approaches, even when applied to the *same* corpus, so cross-corpora comparisons must be carefully considered. There are issues of the DA label set used, the labels considered and those ignored, the pre-processing of the corpus, the use of orthographic information, or prosody and so on. What seems clear is that there are no obvious leading contender for algorithm best suited to the DA classification task. Instead, we focus on the features used for DA classification.

3 Automatic Cue Extraction

When examining prior approaches, we noticed that they used a range of different features for the DA classification task, including lexical, syntactic, prosodic and dialogue context features. Most classifiers used some lexical features (the words in the utterances under consideration), frequently employing some kind of Hidden Markov Modelling to every utterance (Levin et al., 2003; Stolcke et al., 2000; Reithinger and Klesen, 1997), a technique popular in speech processing. We were inspired by the work of Samuel et al. (1999), who instead of modelling entire utterances, extract significant *cue phrases* from the VerbMobil corpus of dialogues. We use a method for cue extraction unused by Samuel et al. (1999).

What defines a good cue phrase? We are looking for words or phrases in a corpus that regularly co-occur with individual dialogue acts. We use

the term *predictivity* to indicate how predictive a phrase is of a particular DA. We want to select phrases that are highly indicative, and so concern ourselves with the highest predictivity of a particular cue phrase. We call this score the maximal predictivity. There are several other thresholds that should also be apparent. First, below some maximal predictivity score, we assume that phrases will no longer be discriminative enough to be useful for labelling DAs. Second, the number of occurrences of each phrase in the corpus as a whole is important. In their experiments, Samuel et al. (1999) constructed all n-grams of lengths 1 through 3 from the corpus, and then applied a range of measures which pruned the n-gram list until only candidate cue phrases remained. In order to test the effectiveness of these automatically acquired cue phrases, Samuel et al. (1999) passed them as features to a machine learning method, in their case transformation-based learning.

More formally, we can describe our criteria, predictivity, for selecting cue phrases from the set of all possible cue phrases in the following way. The predictivity of phrase c for DA d is the conditional probability $P(d|c)$, where:

$$P(d|c) = \frac{\#(c\&d)}{\#(c)}$$

We represent the set of all *possible* cue phrases (all n-grams length 1–4 from the corpus) as C, so given $c \in C$: c represents some possible cue phrase. Similarly, D is the set of all dialogue act labels, and $d \in D$: d represents some dialogue act label. Therefore $\#(c)$ is the count of (possible) cue phrase c in corpus, and $\#(c\&d)$ is the count of occurrences of phrase c in utterances with dialogue act d in the training data. The *maximal predictivity* of a cue phrase c, written as $mp(c)$, is defined as:

$$mp(c) = \max_{d \in D} P(d|c)$$

In their experiments, Samuel et al. (1999) also experimented with conditional probability, using $P(c|d)$, or the probability of some phrase occurring given some Dialogue Act. For our experiments, the word n-grams used as potential cue phrases during are automatically extracted from training data. All word n-grams of length 1–4 within the data are considered as candidates. The maximal predictivity of each cue phrase can be computed directly from the corpus. We can use this value as one threshold for pruning potential cue phrases from our model. Removing n-grams below some predictivity threshold will improve the compactness of the model produced. Another reasonable threshold would appear to be the frequency count of each potential cue phrase. Phrases which have a low frequency score are likely to have very high predictivity scores, possibly skewing the model as a whole. For example, any potential cue phrase which occurs only once will de-facto have a 100% predictivity score. We can use a minimal count value ($t_\#$) and minimal predictivity thresholds (t_{mp}) to prune the set C^* of 'useful' cue phrases derived from the training data, as defined by:

$$C^* = \{c \in C \mid mp(c) \geq t_{mp} \wedge \#(c) \geq t_\#\}$$

The n-grams that remain after this thresholding process are those we identify as cue phrases. For our initial experiments, we used a predictivity of 30% and a frequency of 2 as our thresholds for cue extraction.

4 Cue-Based DA Classification

Having defined our mechanism to extract cue phrases from a corpus, we need some way to evaluate their effectiveness. Samuel et al. (1999) passed their cue phrases as a feature to a machine learning method. We chose instead a method where the cue phrases extracted from a corpus could be used *directly* as a method of classification. If our extracted cues are indeed reliable predictors of dialogue acts, then a classifier that uses these cues directly should perform reasonably well. If, on the other hand, this mechanism did not work, it would not necessarily mean that our cue phrases are not effective, only that we need to pass them to a subsequent machine learning process as others had done. The benefit of our direct classification approach is that it is very fast to evaluate, and gives us immediate feedback as to the possible effectiveness of our automatically extracted cue phrases.

The predictivity of a cue phrase can be exploited directly in a simple model of Dialogue Act classification. We can extract potential cue phrases as described in Section 3. The resulting cue phrases selected using our measure of predictivity are then used directly to classify unseen utterances in the following manner. We identify all the potential cue phrases a target utterance contains, and determine which has the highest predictivity of some dialogue act category, then assign that category. Given the notation we define earlier, we can obtain the DA predicted by a particular cue $(dp(c))$ by:

$$dp(c) = \operatorname*{argmax}_{d \in D} P(d|c)$$

If multiple cue phrases share the same maximal predictivity, but predict different categories, we select the DA category for the phrase which has the higher number of occurrences (that is, the n-gram with the highest frequency). If the combination of predictivity and occurrence count is insufficient to determine a single DA, then a random choice is made amongst the remaining candidate DAs. If $ng(u)$ defines the set of ngrams of length 1..4 in utterance u, and C_u^* is the set of n-grams in the utterance u that are also in the threshold model C^* then C_u^* is defined as:

$$C_u^* = ng(u) \cap C^*$$

Given our thresholds, the $mpu(u)$ (the utterance maximal prediction, or mp value for the highest scoring cue in utterance u) is defined as:

$$mpu(u) = \max_{c \in C_u^*} mp(c)$$

The maximally predictive cues of an utterance $(mpcu(u))$ are:

$$mpcu(u) = \{c \in C_u^* \mid mp(c) = mpu(u)\}$$

Then the maximal cue of utterance $(mcu(u))$, i.e. one of its maximally predictive cues that has a maximal count (from within that set), is:

$$mcu(u) = \operatorname*{argmax}_{c \in mpcu(u)} \#(c)$$

Finally, for our classification model, $dpu(u)$ utterance DA prediction — the DA predicted by model for utterance u, is defined as:

$$dpu(u) = dp(mcu(u))$$

If no cue phrases are present in the utterance under consideration, then a default tag is assigned.

To this basic model, we added three further elaborations. The first used models sensitive to utterance length. When examining the ICSI-MRDA corpus, Ji and Bilmes (2006) found that the mean length of <STATEMENT> utterances was 8.60 words, <BACKCHANNEL> utterances were 1.04 words, <PLACE-HOLDERS> utterances were 1.31 words and <QUESTIONS> utterances were 6.50 words. Taking this as a start point, we grouped utterances into those of length 1 (i.e. short, or one word utterances), those with lengths 2–4 (we call medium length utterances), and those of length 5+ (the long length model, that comprises everything else), and produced separate cue-based models for each group.

Second, we introduced <*start*> and <*finish*> tags to each utterance (independent of the calculation of utterance length), to capture position specific information for particular cues. For example "<*start*> *okay*" identifies the occurrence of the word *'okay'* as the first word in the utterance. Finally, in the Switchboard annotation, there are other markers dealing with various linguistic issues, as outlined in Meteer (1995). A primary example is the label <+>, which indicated the presence of overlapping speech. One approach to better utilise this data is to 'reconnect' the divided utterances, i.e. appending any utterance assigned tag <+> to the last utterance by the *same* speaker. We base the selection of these model elaborations and the values for the parameters of frequency and predictivity on prior research (cf. (Webb et al., 2005a; Webb et al., 2005b; Webb et al., 2005c)).

5 Cue-Based Classification Results

Ultimately, we want to compare classification performance of a set of automatically extracted cue phrases across the two corpora, Switchboard and ICSI-MRDA. Both are annotated with similar variants of the DAMSL annotation scheme,

Condition	Cue Source	Cue Count	Accuracy
(1)	Switchboard training data	136,942	80.72%
(2)	ICSI-MRDA training data	48,856	70.78%
(3)	Intersection of Switchboard and ICSI-MRDA Training Data	25,053	72.34%
(4)	As above, discard <STATEMENT> cue phrases	577	72.62%
(5)	As above, retain only cue phrases containing <*start*> tags	242	72.52%
(6)	As above, retain only cue phrases appearing in every training intersection	148	72.09%

Table 1: Switchboard Classification Results

but there are differences. For example, the ICSI-MRDA corpus introduces several new labels that do not exist in the Switchboard annotation. Some labels in the Switchboard annotation are clustered into a single corresponding label in the ICSI-MRDA corpus, such as the two labels from Switchboard, <STATEMENT-OPINION> and <STATEMENT-NON-OPINION>, which are represented by a single label <STATEMENT> in the ICSI-MRDA corpus. To facilitate cross-corpus classification, we will cluster these labels as described in Shriberg et al. (2004). Of course, any clustering of labels has an impact on classifier performance, usually resulting in an increase. Webb et al. (2005c) indicate that clustering statement labels in the Switchboard corpus should improve performance by 8-10% percentage points.

5.1 Baseline Results

We need to establish baseline classification performance for both corpora. Our baseline for this classification task is to the most frequently occurring label for all utterances. For a number of dialogue corpora, the most frequently occurring label is some sort of statement or assertion, which is true for both the Switchboard and ICSI-MRDA corpora, where <STATEMENT> is the most frequent label. For the Switchboard corpus, selecting this label results in 51.05% accuracy. Remember that we are working with a version of the Switchboard corpus where we have clustered the original labels <STATEMENT-OPINION> and <STATEMENT-NON-OPINION> into a single label. In the original Switchboard annotation, the most frequently occurring label is <STATEMENT-NON-OPINION>, which occurs 36% of the time. Further analysis on the Switchboard corpus by Webb et al. (2005c) high-

lights that a significant number of <STATEMENT-OPINION> utterances in Switchboard are mislabelled as <STATEMENT-NON-OPINION> by human annotators. For the ICSI-MRDA corpus, an accuracy of 31.77% is achieved by labelling each utterance as <STATEMENT>.

Now we have established a simple baseline of performance, we want to know how well our cue-based classification method works applied to these corpora, as an evaluation of how well our cue extraction method works for each of these corpora. We ran a 10-fold stratified cross-validation exercise (referred to as Condition (1) in Tables 1 and 2) using the cue-based extraction mechanism described in Section 3, selecting cue phrases from the training data (which averaged 180k utterances for Switchboard, and 95k utterances for ICSI-MRDA), resulting in an average of 135k cue phrases from Switchboard and 50k cue phrases from ICSI-MRDA. We then applied these cue-based models to the held out test data as described in Section 4, applying Switchboard extracted cue phrases to Switchboard test data, and likewise with the ICSI-MRDA data. This establishes the best performance by our algorithm over these data sets. For Switchboard, we achieve 80.72% accuracy, as predicted by the work reported in Webb et al. (2005c). For ICSI-MRDA we obtain an accuracy of 58.14%. Remember, this model is applied to the 55-label annotated ICSI-MRDA corpus. Best reported classification accuracy for this corpus is the 66% reported by Ji and Bilmes (2005), using a graph-based model that models both utterances and sequences of DA labels. For both corpora, the cue-based model of classification outperforms the baseline, using no dialogue context whatsoever.

Condition	Cue Source	Cue Count	Accuracy
(1)	ICSI-MRDA training data	48,856	58.14%
(2)	Switchboard training data	136,942	47.07%
(3)	Intersection of Switchboard and ICSI-MRDA Training Data	25,053	47.86%
(4)	As above, discard <STATEMENT> cue phrases	577	48.05%
(5)	As above, retain only cue phrases containing <start> tags	242	47.30%
(6)	As above, retain only cue phrases appearing in every training intersection	148	46.34%

Table 2: ICSI-MRDA Classification Results

5.2 Cross-Corpus Results

The focus of our effort is not to maximise raw performance over individual corpora, but to examine the effectiveness of our automatically extracted cue phrases, and one mechanism to do this is to compare classification *cross-corpora*. If our cue phrases are sufficiently general predictors of DA labels across corpora, we believe that to be a powerful claim for cue phrases as a DA classification feature. Therefore, our next step was to take the cue-phrases generated from each fold of the Switchboard experiment, and apply them to the held out test data from the corresponding fold of the ICSI-MRDA experiment, and vice-versa. This is a test to see how generally applicable are the cue phrases extracted from each corpus.

When we take cues extracted from the Switchboard corpus, and apply them to the held out portion of the ICSI-MRDA corpus, we achieve an average classification accuracy (over our 10-folds) of 47.07%. This score represents 81% of the accuracy achieved by our prior result when ICSI-MRDA test data is classified using ICSI-MRDA training data. It also represents 71% of the best published score on this corpus (Ji and Bilmes, 2005). When we classify held out Switchboard test data with cue phrases extracted from the ICSI-MRDA corpus, we achieve an average classification accuracy of 70.78%, which corresponds to 88% of our best score on this corpus using Switchboard training data. These results correspond to Condition (2) in Tables 1 and 2.

These are very positive results for both directions of classification, indicating that the cue phrases we automatically extract from our corpora are generally applicable as a feature for DA classification.

5.3 Cue Phrase Reduction

We have successfully shown that we can use cue phrases extracted from one corpus to classify utterances from a different corpus. We used an inclusive approach, using all cue phrases extracted from the source corpus training data. Intuitively however, we might expect to get comparable performance by using only those cue phrases that appear in both corpora. For these intersection cue phrases, we require a strict overlap. Once the cues phrases are extracted from each individual training fold for each corpus, they are compared and retained if and only if:

- the cue phrase itself is a direct match, including any position specific label

- the DA the phrases predicts is a match

- the model number (as defined in Section 4) is a match

For each fold of our cross-validation, we take cues phrases extracted from the training data that appear in *both* corpora, pruning out cue phrases that only appear in one of the corpora. We then retain *only* those cue phrases that meet these criteria from both corpora for each specific fold, and apply them to the held out test data from that fold for each corpus.

Average classification performance for both corpora rises very slightly in comparison to using all extracted cue phrases. These results can be seen as Condition (3) in Tables 1 and 2. When applying the intersection cue phrases to the Switchboard test data, we achieve an average classification accuracy score of 72.34%. When we apply the intersection cues to the ICSI-MRDA test data, the average score is 47.86%. The average number of cue phrases that are used in this experiment

1315

(i.e that appear in all training folds for both corpora, with matching model information) is around 25k. This represents 50% of the average number of cues extracted from the ICSI-MRDA corpus, and only 19% of the average number of cue phrases extracted from the Switchboard corpus.

We describe earlier that our default label as applied by our classifier when no cue phrase can be found is the <STATEMENT> label, the most frequent single label in both corpora. Given this, we can safely remove cue phrases that predict <STATEMENT> labels from our cue phrase set. The absence of such cue phrases should have no impact on our classification performance, but should reduce our total number of cue phrases. As can be seen in Condition (4) in Tables 1 and 2, this is indeed the case, with no statistical significance between the results with and without <STATEMENT> cue phrases. However, there is a drop in the number of cue phrases. When we remove all <STATEMENT> cue phrases from the intersection of cue phrases, we are left with an average of 577 cue phrases.

Further analysis of classifier performance indicates that a high percentage of actual labelling is performed using a subset of even the cue phrases extracted under Condition (4). We observed that cue phrases that contain a *<start>* tag (as described in Section 4) were used in the majority of cases. Our final experiment was to extract, from the 577 cue phrases, only those phrases that contain the *<start>* tag. This reduced the average number of cue phrases to 242. Classification performance remains unaffected, scoring an average of 72.52% for Switchboard and 47.30% for ICSI-MRDA, as seen in Condition (5) in the results tables. We note that of those 242 phrases, 148 appear in the intersection of *every* training fold of our 10-fold cross-validation. When we use only those 148 cue phrases for classification, as seen in Condition (6), average classification accuracy remains the same; 72.09% for Switchboard, and 46.34% for ICSI-MRDA.

6   Conclusions

In this paper, we investigate a cue-based approach to DA classification, applied to two corpora, Switchboard and ICSI-MRDA. We automat-

ically extracted cue phrases from both corpora, and used them directly to classify unseen utterances from the corresponding corpus, demonstrating that our automatically discovered cue phrases are a sufficiently useful feature for this task.

We then explored the generality of our cue phrases, by applying them directly as a classifier to data from the alternate corpus. Whilst there was some expected drop in performance, the classification accuracy for both experiments is good, given such a small number of features and the simple design of the classifier. The result indicates that cue phrases are a highly useful feature for DA classification, and can be used to classify data from new corpora, possibly as some part of some quasi-automatic first annotation effort.

We experimented with reducing the set of cue phrases, using increasingly restrictive measures of retaining our automatically discovered cues. We found that we did not have a drop in performance compared to the cross-corpus classification accuracy, even when the cue set is drastically reduced (to 0.001% of the original Switchboard cue phrases, and 0.003% of the ICSI-MRDA cue phrases). This appears to be a strong indicator of the discriminative power of some small number of automatically discovered core cue phrases.

References

Ang, J., Y. Liu, and E. Shriberg. 2005. Automatic Dialog Act Segmentation and Classification in Multiparty Meetings. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 1061–1064, Philadelphia.

Bunt, H. 1994. Context and Dialogue Control. *THINK*, 3:19–31.

Core, M., M. Ishizaki, J. Moore, and C. Nakatani. 1999. The Report of the Third Workshop of the Discourse Resource Initiative. *Chiba University and Kazusa Academia Hall*.

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

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

Ji, G. and J. Bilmes. 2005. Dialog Act Tagging Using Graphical Models. In *Proceedings of the IEEE*

International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA.

Ji, G. and J. Bilmes. 2006. Backoff Model Training using Partially Observed Data: Application to Dialog Act Tagging. In *Proceedings of the Human Language Technology/ American chapter of the Association for Computational Linguistics(HLT/NAACL'06)*.

Jurafsky, D., 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.*

Lendvai, P. and J. Geertzen. 2007. Token-Based Chunking of Turn-Internal Dialogue Act Sequences. In *Proceedings of the 8th SIGDIAL Workshop on Discourse and Dialogue*, pages 174–181, Antwerp, Belgium.

Levin, L., C. Langley, A. Lavie, D. Gates, and D. Wallace. 2003. Domain Specific Speech Acts for Spoken Language Translation. In *Proceedings of 4th SIGdial Workshop on Discourse and Dialogue*.

Meteer, M. 1995. Dysfluency Annotation Stylebook for the Switchboard Corpus. Working paper, Linguistic Data Consortium.

Reithinger, N. and M. Klesen. 1997. Dialogue Act Classification Using Language Models. In *Proceedings of EuroSpeech-97*.

Rotaru, M. 2002. Dialog Act Tagging using Memory-Based Learning. Term project, University of Pittsburgh.

Samuel, K., S. 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*.

Shriberg, E., R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Special Interest Group on Discourse and Dialogue (SIGdial)*, Boston, USA.

Stolcke, A., 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*.

Verbree, D., R. Rienks, and D. Heylen. 2006. Dialogue Act Tagging using Smart Feature Selection; Results on Multiple Corpora. *Spoken Language Technology Workshop, 2006. IEEE*, pages 70–73.

Webb, N. and T. Liu. 2008. Investigating the Portability of Corpus-Derived Cue Phrases for Dialogue Act Classification. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, Manchester, United Kingdom.

Webb, N., M. Hepple, and Y. Wilks. 2005a. Dialogue Act Classification Based on Intra-Utterance Features. In *Proceedings of the AAAI Workshop on Spoken Language Understanding*, at the Twentieth National Conference on Artificial Intelligence, Pittsburgh, PA.

Webb, N., M. Hepple, and Y. Wilks. 2005b. Empirical Determination of Thresholds for Optimal Dialogue Act Classification. In *Proceedings of the Ninth Workshop on the Semantics and Pragmatics of Dialogue*.

Webb, N., M. Hepple, and Y. Wilks. 2005c. Error Analysis of Dialogue Act Classification. In *Proceedings of the 8th International Conference on Text, Speech and Dialogue*, Carlsbad, Czech Republic.