Cross-Domain Dialogue Act Tagging

Nick Webb, Ting Liu, Mark Hepple, Yorick Wilks

University at Albany, SUNY, USA
University of Sheffield, UK
{nwebb;tl7612}@albany.edu, {m.hepple;y.wilks}@sheffield.ac.uk

Abstract

We present recent work in the area of Cross-Domain Dialogue Act (DA) tagging. We have previously reported on the use of a simple dialogue act classifier based on purely *intra-utterance* features — principally involving word n-gram cue phrases automatically generated from a training corpus. Such a classifier performs surprisingly well, rivalling scores obtained using far more sophisticated language modelling techniques. In this paper, we apply these automatically extracted cues to a new annotated corpus, to determine the portability and generality of the cues we learn.

1. Introduction

We present our current work in the area of Dialogue Act classification - finding a single label that corresponds to the intention of the user when performing an utterance. For example, the utterance *"Hello"* is a form of greeting, whereas the utterance *"What's your name?"* is a type of question.

A number of researchers (Hirschberg and Litman, 1993; Grosz and Sidner, 1986) speak of cue phrases in utterances that can serve as useful indicators of discourse structure. We have investigated the use of cue phrases to predict dialogue acts (functional tags which represents the communicative intentions behind each user utterance) (Webb et al., 2005a). We developed an approach, in common with the work of Samuel et al. (1999), where word n-grams that might serve as potentially useful cue phrases are automatically detected in a corpus. The novelty in our method is the calculation and use of the *predictivity* of a particular word or phrase, where this measure can be exploited directly in a simple model of dialogue act classification. We have previously reported the results of experiments evaluating this approach on the SWITCHBOARD corpus and our results rival the best reported over that data (Stolcke et al., 2000), although our method adopts a significantly less complex algorithm.

An interesting feature of our approach is that a by-product is a ranked list of cue phrases derived from the source corpus. Visual inspection of these cues reveals that, as one might expect, there is a high degree of correlation between phrases such as *"can you"* and the Dialogue Act <yes/no question>, *"where is"* and *"who is"* with the Dialogue Act <wh-question> and *"right"* or *"ok"* with Dialogue Act <agree/accept>. These cues appear to be of a general nature, unrelated to the source domain or application, and despite being automatically acquired from one domain specific corpus should be applicable to new corpora. It is this hypothesis we wish to test.

This paper presents our work on dialogue act classification using cues automatically extracted from a corpus from one domain, and applying these cues as a classifier over a new corpus from a different domain. The material is presented as follows: Previous work with dialogue act modelling is outlined in Section 2. An overview of the corpora used for this task can be seen in Section 3. We briefly describe our approach to DA classification in Section 4. Our experiments evaluating the cue-based dialogue act classifier tagging new, out-of-domain data are given in Section 5. Finally we end with some discussion and an outline of intended further work.

2. Related Work

Dialogue Acts (DAs) (Bunt, 1994), also known as speech acts or dialogue moves (Power, 1979), represent the functional performance of a speaker's utterance. Searle (1969) is probably most often associated with *speech acts*, building on prior work on illocutionary acts by Austin (1962), as a fundamental concept of linguistic pragmatics, analysing for example what it means to give a greeting such as *"Hello there"*, ask a question like . *"How is your mother?"* or make a request such as *"Can you move your foot?"*. While speech acts provide a useful characterisation of one kind of pragmatic force, more recent work, closely linked to the development of spoken language dialogue systems, has focused on the some of the more conversational roles such acts can perform, and their use in the automatic interpretation of user utterances in spoken language dialogue systems (cf. (Hardy et al., 2005; Pellom et al., 2001; Peckham, 1991)).

There are two broad categories of computational model used to interpret these acts. The first, including the work of Cohen and Perrault (1979) and the TRAINS dialogue system (Allen et al., 1995), relies on processing belief logics, centring on the impact each utterance has on the hearer - what the hearer believes the speaker intended to communicate. These models can be very accurate, particularly at processing indirect cues - utterances whose surface form indicates one act, but in actuality represent another act altogether. Indirect speech acts are commonly used to reject proposals and to make requests. For example a speaker asks, *"Would you like to meet me for coffee?"* and another replies, *"I have class."* The second speaker used an indirect speech act to reject the proposal - the literal meaning of *"I have class"* does not entail any sort of rejection, although the meaning is clear. However, models based on this level of interpretation are often complex, and require significant world-knowledge to create.

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

The second model type is cue-based, and centres on the notion of repeated, predictive cues - subsections of language which are strong indicators of specific DAs. In this second category, much of the work is cast as a probabilistic classification task, solved by training approaches on labelled examples of speech acts. In recent years, a range of state-of-the-art machine learning algorithms have been applied to the leading annotated corpora for this task, including Memory-Based Learning(Fernndez et al., 2004), Graph models (Ji and Bilmes, 2006) and Support Vector Machines (Liu, 2006), with little significant performance difference between these models.

As an example of these probabilistic methods, Reithinger and Klesen (1997) applied a HMM approach to the DA sequences of 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 that exploits both the order of words *within* utterances and the order 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 71.0% 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, although comparison across corpora can be difficult, given different dialogue act taxonomies and overall vocabulary sizes.

Another learning approach applied to dialogue act modelling by Samuel et al. (1998) use 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, of particular relevance to our work, is the automatic identification of word sequences that might serve as useful dialogue act cues (Samuel et al., 1999). A number of statistical criteria are applied to identify potentially useful word n-grams that are then supplied to the transformation-based learning method as 'features'. What has not been explored is the portability or generality of the models that are acquired using these methods.

3. Experimental Corpora

Our work as described here applies to two corpora - the DA-tagged portion of the SWITCHBOARD corpus (Jurafsky et al., 1998), and the AMITIÉS GE corpus (Hardy et al., 2002; Hardy et al., 2003), created as part of the AMITIÉS European 5th Framework program project (Hardy et al., 2005). A summary of the two corpora can be seen in Figure 2.

3.1. Switchboard

The annotated portion of the SWITCHBOARD corpus comprises 1155 annotated conversations between two human participants, where the dialogues are of an unstructured, non-directed character. Participants do not know each other, and are provided only with a set of topics which they may wish to discuss. The SWITCHBOARD corpus is annotated using an elaboration of the DAMSL tag set. In 1998 the Discourse Resource Initiative finalised a task-independent set of DAs, called DAMSL (Dialogue Act Markup in Several Layers), for use across different domains. DAMSL has been used to mark-up several dialogue corpora, such as TRAINS (Core and Allen, 1997), and the SWITCHBOARD corpus (Jurafsky et al., 1998). DAMSL draws both on the need

Corpus	*Availability*	*Utterance count*	*Dialogue count*	*Word count*	*Distinct words*	*Dialogue type*
SWITCHBOARD	public	223606	1155	1431725	21715	Conversational
AMITIÉS GE	restricted	30206	1000	228165	7841	Task-oriented

Figure 2: Summary data for the dialogue corpora

to provide a reliable corpus to derive cue models, and the philisophical underpinnings from earlier work. For example, DAMSL includes a series of forward-looking functions, such as <influence-on-addressee>, similar to Searle's directives, and <influence-on-speaker>, which represents the commissives.

The annotation over the SWITCHBOARD corpus involves 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 these 220 tags into 42 larger classes, listed in Figure 1, and it is this clustered set that was used both in our experiments and those of Stolcke et al. (2000). In measuring the agreement between annotators in labelling this data, Jurafsky et al. (1998) report an average pair-wise kappa of .80 (Carletta, 1996). An excerpt of dialogue from the SWITCHBOARD corpus can be seen in Figure 3.

3.2. AMITIÉS

The AMITIÉS project (Hardy et al., 2005) collected 1000 English human-human dialogues from European GE call centres. These 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 resulting data has been sanitised, to replace identifying features such as names, addresses and account numbers with generic information (*"John Doe"*, *"1 The Street"*) and the corpus is annotated with DAs using XDML, combining a variant of DAMSL (Hardy et al., 2002) with domain specific semantic information such as account numbers and credit card details (Hardy et al., 2003).

The most frequent tag in the AMITIÉS corpus is *Influence-on-listener="Information-request"*, which occurs 20% of the time. For this corpus, the average pair-wise kappa score of .59 was significantly lower than the SWITCHBOARD corpus. For the major categories (questions, answers), average pair-wise kappa scores were around .70, indicating a higher degree of consistency between annotators for these major classes. Again, according to the work of Carletta (1996), a minimum kappa score of 0.67 is required to draw tentative conclusions. An excerpt of dialogue from the AMITIÉS corpus can be seen in Figure 5.

4. DA Classification

In this section we describe our approach to DA classification, based solely on intra-utterance features. A key aspect of the approach is the selection of the word n-grams to use as cue phrases. 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. (1999) do not consider. 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. 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: $P(d\,|\,n)$, a value that 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. This thresholding achieves two benefits. The size of the stored model is reduced, but most importantly it removes some low frequency, high predictivity n-grams that skew classification performance. The n-grams that remain are used as cue phrases. The threshold values that are used in all experiments were arrived at empirically, using a validation set to automatically set the threshold levels independently of the test data, as described in Webb et al. (2005b). Ideally, such thresholding finds the balance between removing cue phrases, and optimal performance.

4.2. Using Cue Phrases in Classification

To classify an utterance, we identify all the word n-grams it contains, and determine which of these has the highest predictivity of some dialogue act category (i.e. is performing as some cue). If multiple cue phrases share the same maximal predictivity, but predict different categories, we select the DA for the phrase which has the higher number of occurrences. If the combination of predicitivity and occurrence count is insufficient to determine a single DA, then a random choice is made amongst the remaining candidate DAs. If no cue phrases are present, then a default tag is assigned, corresponding to the most frequent tag within the training corpus.

In prior work we performed five different experiments using a variety of simple methods for pre-processing the data

<Turn Id="utt1" Speaker="A" DA-Type="Hold-before-answer"> **uh, let's see**</Turn>

<Turn Id="utt2" Speaker="A" DA-Type="Abandoned"> **How about ten years ago,**</Turn>

<Turn Id="utt3" Speaker="A" DA-Type="Open-question"> **what do you think was different ten years ago from now?**</Turn>

<Turn Id="utt4" Speaker="B" DA-Type="Statement-opinion"> **Well I would say as far as social changes go I think families were more together.**</Turn>

<Turn Id="utt5" Speaker="B" DA-Type="Statement-opinion"> **They did more things together**</Turn>

<Turn Id="utt6" Speaker="A" DA-Type="Acknowledge"> **Uh-huh**</Turn>

Figure 3: Excerpt of dialogue from the SWITCHBOARD corpus

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

Figure 4: Example SWITCHBOARD utterance incorrectly labelled

(Webb et al., 2005a). Our best reported figures on the 202k utterance SWITCHBOARD 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). During these experiments we observed that some dialogue act categories seemed to be most easily confused - where utterances of one category are consistently incorrectly tagged as being of a second category - a view confirmed in a subsequent error analysis (Webb et al., 2005c). For the most part, these errors fall into two categories - poor annotation of the data, where two categories have been inconsistently assigned (as with <statement-opinion> *vs.* <statement-non-opinion>, a common example of which can be seen in Figure 4), and those DA's that have consistently similar lexicalization, such as <agree/accept> and <backchannel>, which are both often realised with *"yeah"* or *"ok"*. For the first kind there is little mechanically than can be done (collapsing the two categories together results in an immediate 10% point increase in tagging accuracy, which approaches the upper bound of performance as indicated in the inter-annotator agreement scores), and for the latter, this is a problem that this intra-utterace technique cannot resolve. We also presented information that shows that adding a sequence model of DA progressions - an n-gram model of DAs - results in no increase in performance (Webb et al., 2005a). This is surprising considering that Stolcke et al. (2000) report their best figures when *combining* a HMM model of the words inside utterances with a tri-gram model of the Dialogue Act sequence, as in the work of Reithinger and Klesen (1997). Adding the sequence model to the HMM language model adds around 20% points to the final accuracy score over the SWITCHBOARD data. On the basis of this result, we hypothesise that our cues are highly predictive of dialogue structure, and that much dialogue processing can take place at a very shallow level.

5. Cross-Domain Classification

The central purpose of this paper is to examine the use of automatically extracted cues to tag data *other* than the corpus from which they are derived. The hypothesis we wish to test is that these cues are sufficiently general to work as a classification device on a corpus from a different domain, even containing interactions of a different conversational style. Specifically, SWITCHBOARD is an open domain spoken human-human conversational corpus and we have shown state-of-the-art tagging performance over this data using the cue-based model. We now wish to see how well these same cues perform over the AMITIÉS GE corpus of spoken *t*ask-based dialogues. The dialogues in the AMITIÉS GE corpus are far more goal directed, and contain domain specific cues that are not found in the general conversational SWITCHBOARD corpus.

The ability to apply cues extracted from one corpus to new data is an interesting challenge. It could confirm work which indicates the prominence of such word cues in language (Hirschberg and Litman, 1993). The fact that such cues can be general across domains and applications is of obvious interest. A tag mechanism that can operate across domains presents a range of benefits - for example it can be used to annotate or partially annotate new data collections, or such generic mark-up of dialogue function might provide the basis for discovering higher level dialogue structure, such as clarification or error-correction.

5.1. DA Mapping

This classification would be simplified if both corpora were annotated with identical DA taxonomies. In actuality, the SWITCHBOARD corpus and the AMITIÉS GE corpus are both annotated with *variants* of the DAMSL DA annotation scheme. In the SWITCHBOARD corpus, the hierarchical nature of the DAMSL schema has been flattened and clustered, to produce 42 major classes. In the AMITIÉS GE corpus, the dialogue level schema has been left largely untouched from

```
<Turn Id="2.1" Speaker="Operator" Info-level="Communication-mgt"
Conventional="Opening">good morning customer services sam speaking</Turn>

<Turn Id="2.2" Speaker="Operator" Info-level="Task" Forward-function="Offer">
how can i help</Turn>

<Turn Id="3.1" Speaker="Customer" Info-level="Communication-mgt"
Conventional="Opening">erm good morning</Turn>

<Turn Id="3.2" Speaker="Customer" Info-level="Task"
Forward-function="Explanation">erm I was away for about two months and i came back
and my card i don't know whether i have lost it or it is stolen</Turn>

<Turn Id="4.1" Speaker="Operator" Understanding="Backchannel"
Response-to="T3.2">right okay</Turn>

<Turn Id="4.2" Speaker="Operator" Info-level="Task"
Influence-on-listener="Info-request-explicit">can you confirm your name
for me please</Turn>
```

Figure 5: Excerpt of dialogue from the AMITIÉS GE corpus

the DAMSL original. In terms of the hierarchy, we are fortunate in that most cases, levels can be inferred from one another - that is, if we can identify the salient part of the annotation, that informs the other parts of the annotation. However, in order then to be able to compare automatic classification performance across the two corpora, a mapping needs to be created between the 42-class schema of SWITCHBOARD and the DAMSL-like XDML schema of the AMITIÉS GE corpus.

In their work, Jurafsky et al. (1998) include such a mapping between SWITCHBOARD and DAMSL that covers approximately 80% of the labels in the SWITCHBOARD corpus. We have adapted this slightly to cover minor differences between the XDML used in the AMITIÉS GE corpus and the original DAMSL, although this leaves us with two issues that we need to address.

First there are differences in granularity on both sides. Importantly, in many instances we may identify the most salient role of the utterance, but miss modification information which may make little interpretative difference. For example, mark-up in the AMITIÉS GE corpus makes the distinction between <Forward-function="Assert"> and <Forward-function="Reassert">, whereas mark-up in the SWITCHBOARD corpus ignores such a distinction, and accepts only that these are both of type <Forward-function="Assert"> - although the SWITCHBOARD corpus captures the difference between assertions that are opinions, and those that are not. The original DAMSL does not capture this distinction. To address this mismatch we create a set of super classes by relating the annotations of SWITCHBOARD-DAMSL and the AMITIÉS GE-XDML corpora at the most salient level, according to the mapping contained in Jurafsky et al. (1998). Whilst the majority of tags have a one-to-one correlation, there are elements of both the Forward-Looking Function (see Figure 6) and Backward-Looking Function (Figure 7) that require mapping in both directions.

Secondly, there are a number of AMITIÉS GE tags that we know a-priori we have little or no chance to recognise. For example, the AMITIÉS GE corpus is meticulously annotated to include that certain utterances are perceived as answers to prior utterances (such as the fifth utterance in Figure 5. Our approach to DA tagging is purely *intra*-utterance, taking no account of the wider discourse structure, so will not recognise these distinctions. Although such a model of discourse structure should be trivial, based for example on an adjacency pair approach, we have shown that the inclusion of such a model makes no impact on classification performance, although this will be evaluated further in future work.

These issues require that we create two evaluation criteria for our subsequent experiments - **strict** and **lenient**. With strict evaluation, we are required to match *all* of the AMITIÉS GE corpus annotation - despite knowing in advance that this is not possible for a range of utterances. We use our strict evaluation criteria to establish a lower bound of performance for our classifier. Our lenient approach is used as a back-off model, where we require that we correctly identify the most critical part of the multi-part annotation - those that are identified as the most salient.

We'll use the dialogue excerpt shown in Figure 5 as an example of how these two scoring mechanisms work. The first utterance is marked as <Info-level="Communication-mgt" Conventional="Opening">. This has a one-to-one correlation with the SWITCHBOARD-DAMSL tag <conventional-opening>. In the case of this example, and all instances in the AMITIÉS GE corpus, utterances are of <Info-level="Task">, *unless* they are from a small set of exceptions, including openings, closings or backchannels, that are annotated as <Info-level="Communication-mgt">. Once an utterance is tagged as one of these exceptions, we know to change the Info-level accordingly. There

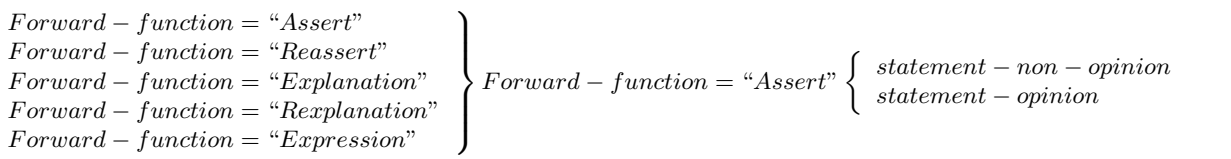

Figure 6: Partial Forward-Looking Function mapping table (XDML } SUPERCLASS { SWITCHBOARD-DAMSL)

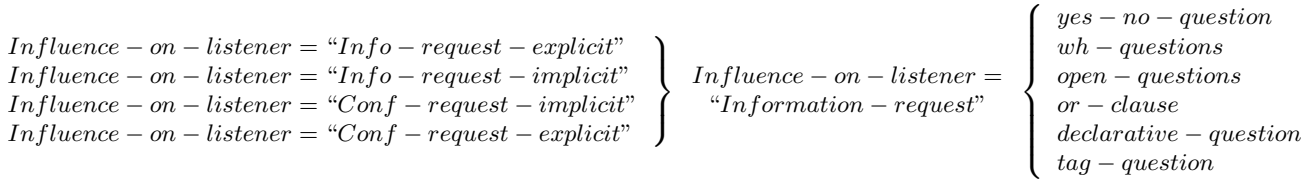

Figure 7: Partial Backward-Looking Function mapping table (XDML } SUPERCLASS { SWITCHBOARD-DAMSL)

is no difference between our strict and lenient evaluation models for the interpretation of this utterance. The same is true for the second, third and fifth utterance annotations, all of which have direct correlation with SWITCHBOARD-DAMSL annotations. However, the fifth utterance includes a <Response-to="T3.2"> annotation, that we will not identify using our intra-utterance model. This utterance will be judged correct using the lenient model, and incorrect using the strict metric.

The fourth utterance is marked as <Forward-function="Explanation">. Using the Forward-function map shown in Figure 6, we see that this maps to the super class <Forward-function="Assert">, that in turn maps to the SWITCHBOARD-DAMSL tags <statement-non-opinion> and <statement-opinion>.

When the cue-phrase classifier is trained over the SWITCHBOARD corpus and applied to the AMITIÉS GE corpus, this means that any utterance that is identified by the presence of a cue phrase as either <statement-non-opinion> or <statement-opinion> will in fact be tagged as <Forward-function="Assert">. Whilst this annotation captures the salient behaviour of the utterance, it is not an exact match to the original AMITIÉS GE corpus annotation. Correspondingly, when scoring, the lenient model will score this as correct, whereas the exact model will not.

The same is true with the sixth utterance, annotated in this case as <Influence-on-listener="Info-request-explicit">. A classifier trained over the SWITCHBOARD corpus would identify this (through the mapping see in Figure 7) as <Influence-on-listener="Information-request">, which would be scored as correct using the lenient measure, and incorrect using the exact.

5.2. Experiments

We first establish our baseline tagging performance. To that end, we take the classification algorithm outlined earlier in Section 4, and apply it to the SWITCHBOARD corpus for both training and testing, replicating the work reported in Webb et al. (2005a). In this case, 198,000 utterances are used for training, and a separate 4,000 utterances are used for testing.

Secondly, we repeat this experiment, substituting the AMITIÉS GE corpus for the SWITCHBOARD corpus in both steps - training and testing. This should give us an upper bound of performance of this particular classification algorithm over this data. In this experiment, we used 10% of the corpus for testing - giving us a total of 27,000 utterances for training and 3,000 utterances for testing.

On a related note, part of the work conducted in Webb et al. (2005a) studied the impact of different size training models when classifying SWITCHBOARD data, using models of 4k, 50k and 202k utterances. Whilst substantial improvement was seen when moving from 4k utterances to 50k utterances, the subsequent increase to 202k utterances had a negligible increase in classification accuracy over the 50k model.

Finally, we attempt cross-domain classification: First, we train our classifier using SWITCHBOARD data, and test using AMITIÉS GE data. Then we apply the classification in reverse - we train on AMITIÉS GE data, and test on the SWITCHBOARD corpus, using all available data in both cases. For the last experiment, we also study the effect of limiting the training data on cross-domain classification, by reducing the SWITCHBOARD data to match that of the AMITIÉS GE training set - that is, to use only 27,000 utterances of the SWITCHBOARD corpus as training data to extract cues, which are then applied both to itself (for reference), and to the AMITIÉS GE corpus.

For all experiments where AMITIÉS GE data is used as a test corpus, both strict and lenient scoring will be used. Strict scoring sets a lower bound for this exercise, and should be greater than chance, which corresponds to the distribution of the most frequent DA tag in each corpus.

5.3. Results

The results of our experiments are summarised in Figure 5. The first experiment, using SWITCHBOARD data for both training and testing, achieves a cross-validated score

Training corpus	*Training utterances*	*Testing corpus*	*Test utterances*	*Common tag (%)*	*Lenient score*	*Strict score*
SWITCHBOARD	198,000	SWITCHBOARD	4,000	36%	n/a	69%
AMITIÉS GE	27,000	AMITIÉS GE	3,000	20%	70.8%	65.9%
SWITCHBOARD	198,000	AMITIÉS GE	30,000	20%	55.7%	39.8%
AMITIÉS GE	27,000	SWITCHBOARD	198,000	36%	48.3%	40%
SWITCHBOARD	27,000	AMITIÉS GE	3,000	20%	53.2%	38%
SWITCHBOARD	27,000	SWITCHBOARD	3,000	36%	n/a	60%

Figure 8: Experimental Results

of 69.6%, where the most frequent tag in SWITCHBOARD, <statement-non-opinion>, occurs 36% of the time. This is a confirmation of the work reported in Webb et al. (2005a), and again demonstrates that this simple model works exceptionally well for this task.

For the second experiment, where we apply this same algorithm to the AMITIÉS GE corpus for both training and testing, we report both lenient and strict scoring. For strict scoring, where we are required to match all the elements of the AMITIÉS GE XDML tag, we score 65.9% accuracy. For lenient, where we must match only the most salient features, we score 70.8% accuracy. Whilst there is no direct comparison to other work on this corpus, Hardy et al. (2005) show partial results for DA classification on this task, looking only at a few major classes, and achieve a score of 86%. However, this includes only the 5 most frequent DA categories, and considers utterances shorter than a certain number of words.

For the cross-domain classification, there are two experimental variants. Initially, we train using all of the SWITCHBOARD data, and test over the complete AMITIÉS GE corpus, and we recorded a strict evaluation score of 39.8% tagging accuracy. Using the lenient score, we achieve around 55.7% accuracy. This can be considered a very good result, given the lower bound score of 20% - that is the count of the most frequent tag.

Then, applying the same experiment in reverse, we train with the AMITIÉS corpus, and test over the SWITCHBOARD data. Using the strict evaluation metric, we achieve a score of 40.0%, and a lenient score of 48.3%. This compares to a baseline of 36%, so is not a drastic improvement over our lower bound. Some inspection of the data informed us that the AMITIÉS GE data did not include many <backchannel> utterances, so subsequently most of these instances in the SWITCHBOARD corpus were missed by our classifier. By changing the default tag to be <backchannel>, rather than the most frequent tag for the training corpus, we achieve a performance gain to 47.7% with strict scoring, and 56.0% with the lenient metric.

In the final experiment, using a size adjusted variant of this cross-domain classification experiment, we score 53.2% with the lenient metric, and 38% with strict, indicating that the reduction is size of the training data has little effect on classification accuracy. Interestingly, in comparison to earlier work, we report a slightly lower value of 60% classification accuracy over the SWITCHBOARD corpus, when training with the reduced set. However, this score can be misleading. Given the size of the original corpus, we are able to perform 7 individual 10-fold cross validation experiments, and we report the average across all 7 of these. Across the 10-fold cross validation experiments, there is a significant variance of results - between an average low of 53.8% and an average high of 67.4%, with a single run high score of 72.5% (compare this to the best single run score reported in Stolcke et al. (2000) of 71%). This figure alone seems to suggest that it is less the size of training data than the composition that is important, but we discuss this further in the following section.

6. Discussion, Future Work

We have shown that the cues extracted from the SWITCHBOARD corpus can be used to classify utterances in a new domain, that of the AMITIÉS GE corpus. We achieve almost 80% of the upper baseline performance over the AMITIÉS GE corpus, when judged using our lenient scoring mechanism - scoring 55.7% using the cross-domain cues, compared to the 70.8% when using in-domain cues. When using the strict measure we still achieve around 60% of the upper bound performance, both results being a substantial improvement over the baseline measure of 20%, corresponding to the most frequent tag. This is a significant result, which confirms the idea that cues can be sufficiently general across domains to be used in classification.

However, whilst the experiment using SWITCHBOARD corpus derived cues to classify AMITIÉS GE data works well, the same is not true in reverse. There are two possible explanations for this result. It could be related to the size of data available for training, although our experiments in this area seem to suggest otherwise. We believe that the composition of the training data is a more crucial element. The fact that SWITCHBOARD corpus data is not domain specific, and, although the DA distribution in this corpus is skewed, it contains enough data for the major classes to be effective on new data. Although the AMITIÉS GE contains a lot of questions and statements, there is very little of the other significant categories, such as <backchannels>, a key DA in the SWITCHBOARD corpus and conversational speech in general. Correspondingly, the cues derived from the AMITIÉS GE data perform well on a selection of utterances in the SWITCHBOARD corpus, but very poorly on others.

We want to perform an in-depth error analysis to see if the errors we obtain in classification accuracy are consistent. We can also compare our list of automatically derived cues phrases, particularly those that overlap between the two corpora, to those reported in prior literature. It might be interesting to see if more complex models, derived

using state-of-the art machine learning approaches, could demonstrate similar portability - i.e is it the simplicity of our model that allows for the observed robust portability? Finally, we wish to combine SWITCHBOARD and AMITIÉS corpora in the cue learning phase, to see how this effects classification, and apply the results to a range of other corpora, including the VERBMOBIL corpus (Reithinger and Klesen, 1997), and the ICSI-MRDA corpus (Shriberg et al., 2004).

7. References

J. Allen, L. Schubert, G. Ferguson, P. Heeman, C. Hwang, T. Kato, M. Light, N. Martin, B. Miller, M. Posesio, and D. Traum. 1995. The TRAINS project: a case study in building a conversational planning agent. *Journal of Experimental and Theoretical Artificial Intelligence*, 7:7–48.

J. L. Austin. 1962. *How to Do Things with Words*. Oxford University Press, Oxford.

Harry Bunt. 1994. Context and dialogue control. *THINK*, 3:19–31.

J. C. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22.

P. R. Cohen and C. R. Perrault. 1979. Elements of a plan based theory of speech acts. *Cognitive Science*, 3.

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.

R. Fernndez, J. Ginzburg, and S. Lappin. 2004. Clarifying ellipsis in dialogue: a machine learning approach. In *Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland*.

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

Hilda Hardy, Kirk Baker, Laurence Devillers, Lori Lamel, Sophie Rosset, Tomek Strzalkowski, Cristian Ursu, and Nick Webb. 2002. Multi-layered dialogue annotation for automated multilingual customer service. In *Proceedings of the ISLE workshop on Dialogue Tagging for Multimodal Human Computer Interaction, Edinburgh*.

H. Hardy, K. Baker, H. Bonneau-Maynard, L. Devillers, S. Rosset, and T. Strzalkowski. 2003. Semantic and dialogic annotation for automated multilingual customer service. In *Eurospeech, Geneva, Switzerland*.

H. Hardy, A. Biermann, R. Bryce Inouye, A. McKenzie, T. Strzalkowski, C. Ursu, N. Webb, and M. Wu. 2005. The AMITIÉS System: Data-Driven Techniques for Automated Dialogue. *Speech Communication*, 48:354–373.

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

Gang Ji and Jeff 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)*.

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.

Yang Liu. 2006. Using SVM and Error-correcting Codes for Multiclass Dialog Act Classification in Meeting Corpus. In *Proceedings of Interspeech (ICSLP)*.

J. Peckham. 1991. Speech understanding and dialogue over the phone: an overview of progress in the sundial project. In *Proceedings of the 2^{nd} European Conference on Speech Communication and Technology*, pages 1469 – 1472.

B. Pellom, W. Ward, J. Hansen, K. Hacioglu, J. Zhang, X. Yu, and S. Pradhan. 2001. University of Colorado Dialog Systems for Travel and Navigation. In *Human Language Technology Conference (HLT-2001)*, San Diego, USA.

Richard J. D. Power. 1979. The organisation of purposeful dialogues. *Linguistics*, 17:107–152.

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*.

J. R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge.

E. Shriberg, 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.

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. 2005a. Dialogue Act Classification Based on Intra-Utterance Features. In *Proceedings of the AAAI Workshop on Spoken Language Understanding*.

Nick Webb, Mark Hepple, and Yorick Wilks. 2005b. Empirical determination of thresholds for optimal dialogue act classification. In *Proceedings of the Ninth Workshop on the Semantics and Pragmatics of Dialogue*.

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