Learning to classify utterances in a task-oriented dialogue

William Black, Paul Thompson, Adam Funk and Andrew Conroy
Department of Computation
UMIST, PO Box 88, Manchester, M60 1QD
{wjb,thompson,jfunk,aconroy}@co.umist.ac.uk

Abstract

In spoken dialogue systems, the identification of the dialogue act of user utterances is an important first step to understanding what the user said, and how the system should respond. In this paper, we present preliminary experimental results of 2 simple approaches to dialogue act classification of user utterances in spoken email dialogues, namely decision tree and Naive Baysian classifiers. Each type of classifier is trained on utterance features, extracted either from the utterance to be classified, or from recently preceding utterances. Transcribed utterances are parsed by a functional dependency grammar, the output of which provides the source for the majority of features used in the trees. Other features include utterance length and, for preceding utterances, the dialogue acts assigned. Preliminary results are encouraging compared to those obtained by more complex methods.

1 Introduction

Spoken dialogue systems, despite the speech-based interaction modality, have a lot in common with text-based NLP applications. Both are complex enough to require a highly modular architecture that integrates a large number of modules dedicated to particular levels or aspects of the analysis. There is, in both, an emphasis on exploiting to the full robust but shallow analyses because of the difficulties of adapting more theoretically motivated approaches to dealing with the variety of usage in a corpus or when an application is deployed. In both cases, some component parts of the analysis are well suited to empirical or inductive development, but others, involving a less "surface level of analysis or planning, are more resistant.

The divergence between theory and practice in dialogue systems is widening as more telephony applications are fielded and the industrial strength of the component technologies improves. Part of the reason for this is the same as in text-based NLP, that pragmatically developed applications are easier to tune for performance at well-defined tasks than software with theoretical pretensions. However, it is also a consequence of the technical limitations imposed by the automatic speech recognition components on which dialogue applications depend. Whilst dictation applications, being attuned to a single speaker, can cope with recognition of unrestricted input, speaker-independent recognition only works effectively when strongly constrained by a grammar of expected inputs.

The industry view is that this limitation in turn enforces the use of system-initiative dialogue management, with very simple dialogue models, intelligible to an average programmer, of which Voice-XML is the latest manifestation. This is a world away from the style of theoretical research, e.g. (Bunt and Black, 2000), in which the starting point is a view of the dialogue participant (including the system) as a rational agent who has to compute complex abductions over extensive world knowledge in order to understand and react.

2 The Athos Architecture and AthosMail

In the Dumas project, a generic architecture for the development of adaptive dialogue management (Athos) is being instantiated by a spoken email reading application (AthosMail). AthosMail should provide richer interaction and more functionality that the current state of the art. Whilst using commercial technology for ASR, the Athos agent-based architecture allows for experimentation with a variety of components for sub-tasks of analysis and planning, and flexible dialogue man-

agement in which the user may take the initiative. The system is being developed in three languages (English, Finnish and Swedish) and partly because of this, analysis involves the collaboration of many distinct agents.

Speech recognition will rely on phrase structure grammars, which are being developed by induction on dialogue corpora, not just on a per-language basis, but also on a per-dialogue state basis, so that the individual grammars can be compact enough to result in acceptable recognition accuracy. Whilst these are being developed using inductive methods, the results are not mature enough to report on in detail at present.

Surface parsing is being done using Functional Dependency Grammar, which is available for all three languages.

Going from surface syntax to interpretation and hence to response is done with multiple agents. Some, like those of commercial toolkits (e.g. Nuance SpeechObjects) are dedicated to particular dialogue states or tasks like requesting a repetition or spelling-out. Others are involved in building logical forms from parse trees, and yet others take short-cuts to analysis, such as determining the speech act type directly from a combination of surface features and dialogue state, the latter evidenced only by previous utterances. The system intentionally uses redundant components, not just so that competing agents can be evaluated, but also because time for analysis is limited, and it may be necessary to react more quickly than a Prolog-technology parser-interpreter can build logically complete representations of user intentions. The rest of this paper describes results attained to date with the empirical development of an agent for dialogue act type recognition.

We expect the results reported below for this preliminary work to be read as work-in-progress. Due to the small size of the corpus and its rather specialized nature, no reliable conclusions can be drawn about the relative merits of the classifiers compared with others reported in the literature. However, some indications about the relative performance of different feature permutations are useful input to the design of an agent for the fist prototype integrated dialogue management system.

3 Dialogue acts and their classification

In order for a spoken dialogue system to respond appropriately to a user utterance, a fundamental task is to determine the dialogue act type (DA) realized by the utterance. For example, the user may have issued a command for some action to be carried out by the system, requested that the system provide some information, or they may be responding to a system question.

The task of automatically classifying utterance DAs is not new one, and several attempts have been reported, using human-human dialogues as well as human-computer dialogues. Probablistic approaches have perhaps been most common; examples include (Reithinger and Klesen, 1997) and (Stolcke et al., 2000). In both approaches, the most likely DA tag for a given utterance is predicted based on the sequence of words in the utterance, combined with some other information. Reithinger and Klesen use information about the dialogue history, whilst Stolcke et al. use an n-gram DA grammar, together with prosodic information about the utterance, modelled using decision trees. (Samuel et al., 1998) use a non-probablistic approach, i.e. transformation-based learning, to produce a set of rules for DA classification. The rules make use of features such as cue phrases found in the utterance, speaker information, number of words in the utterance and the DA tag on the previous utterance.

The decision trees trained by (Stolcke et al., 2000) purely on prosodic utterance information did not prove particularly successful, with a classification accuracy of around 45.4% when applied to an independent test set. The information obtained from the tree could however boost classification accuracy compared with using the probablistic dialogue grammar and word sequence information alone. In this paper, we investigate whether decision trees trained on alternative sets of non-prosodic utterance features can achieve more acceptable results when applied to the problem of classifying user utterances in dialogues with a spoken email system. According to (Mitchell, 1997), Naive Baysian classifiers are competetive with, and may sometimes outperform, decision tree classifiers. Therefore, we compare the results

of the decision tree classifiers with those achieved by Naive Baysian classifiers, trained on the same sets of features.

We use some of the same features as (Samuel et al., 1998), i.e. utterance length and dialogue acts assigned to recently preceding utterances in the dialogue. However, we also make use of structural information about the utterance, obtained from a functional dependency grammar (FDG) parser, developed by Conexor. The parser output makes it possible to extract "important" words from utterances, such as the main verb and its object. To train the classifiers, we used the WEKA system (Witten and Frank, 2000), which provides Java implementations of a number of machine learning algorithms. As each algorithm requires the input data to be of the same format, it is relatively straightforward to experiment with training different types of classifiers.

4 Tagging the Corpus

The corpus was collected using the Wizard Of Oz technique (Dahlbäck et al., 1992). As we do not yet have a working spoken email system, this technique has been used to collect some reasonably authentic human-computer dialogues. In total, we have 18 dialogues, with a total duration of 115 minutes. There are 1356 system utterances and 581 user utterances. The dialogues were transcribed by hand and segmented into utterances. DA tags were then manually assigned to each utterance.

Our tag set consists of 23 tags, which are organised as a hierarchy. The further down the hierarchy, the more specific the tag. It is based partly on the tag sets used for the Map Task (Carletta et al., 1996) and VERBMOBIL (Jekat et al., 1995), with ad-hoc modifications to tailor it to the needs of spoken dialogue systems. Whilst it is preferable for annotators to assign a leaf node tag to each utterance, they are also permitted to assign a more general tag, should the exact purpose of the utterance be unclear.

Under the root node of the hierarchy, ACT, are four branches: UNCLEAR (an unintelligible utterance), REQUEST (an utterance that introduces a goal or subgoal), RESPONSE (an utterance that discharges a goal or subgoal) and PHATIC (an utterance that maintains the communication channel). REQUEST, RESPONSE and PHATIC all have more specific tags beneath them.

5 Utterance Features

Choosing a suitable set of attributes to represent instances is critical to the success of the classifier. We have developed a framework that enables us to define new types of attributes, and to experiment with training classifiers that use different combinations of these attributes. An accompanying graphical user interface allows these tasks to be carried out very simply. It also has facilities to view the trained trees, and to perform evaluations of them by applying the trained tree to an unseen set of test data.

Features extracted from the FDG parser are user-defined, and are specified using regular expressions. This provides a flexible approach to try out a number of different attribute types. In addition, 2 "fixed" features are available, i.e. the number of words in the utterance and the DA assigned. Features may be extracted from the current utterance (i.e. the one to be classified) or from recently preceding utterances. DA tags are obviously only available for previous utterances.

The FDG parser produces both morphological information for word-form tokens and functional dependencies representing structural information within sentences. To illustrate this, Figure 1 shows the output from the parser when given as input the utterance "Check any outstanding messages".

Each line of the output corresponds to a word in the input sentence, and consists of 5 fields: word position, word form, base form, functional dependency and finally a bundle of tags encoding the functional tag, surface syntax tag and morphological tags. The functional dependencies represent the structural information about the sentence. They consist of a function type label and a numerical index that points to its head. In the above utterance, "check" is shown as main element in the sentence. As its head is shown as 0, it is at the root of the dependency tree. It has 1 direct dependent, "messages", which has been identified as its object. In turn, the direct dependents of "messages" are "any" and "outstanding".

```
1  check        check        main:>0    @+FMAINV %VA V IMP
2  any          any          det:>4     @DN> %>N DET
3  outstanding  outstanding  attr:>4    @A> %>N A ABS
4  messages     message      obj:>1     @OBJ %NH N NOM PL
5  <s>          <s>
```

Figure 1: Sample FDG parser output

This parser output can help to identify information about potentially important words in the utterance, such as the main element and its object. For each such word, several types of attribute value may be extracted, including the word itself and the tags assigned to it.

For word-valued attributes, there is an almost infinite number of potential values, whereas the classifiers that we are using require the full set of values for nominal attributes to be specified in advance. However, it is not necessary to try to provide an exhaustive list of values, because only those values that occur frequently in the training data are likely to provide useful evidence for utterance classification. Indeed, values that only occur a small number of times in training data could have a negative effect on the classifier, as the tree may make incorrect classification decisions based on "chance" occurrences of these values. Therefore, the set of values specified for nominal attributes in our framework consists of the n most popular values for the attribute in the training corpus, plus an extra value, "misc" for all other words.

6 Experimental Results

Three sets of experiments have been performed: the first set trained classifiers on only lexical utterance features, the second used only non-lexical features and the third used a combination of lexical and non-lexical features.

Some data from the original corpus has been removed for the purposes of the experiments. Two dialogues have not yet had FDG parser output added to them, whilst in the remaining dialogues, some utterances were found to be incorrectly segmented. As these could hinder the training process, they have been omitted. The remaining 16 dialogues, containing a total of 502 user utterances, were used to train and test the classifiers. Due to the small amount of data available, the testing has been carried out using 10-fold cross validation over the entire corpus. As a baseline accuracy for the performance of the classifiers, we take the proportion of the most popular user DA tag in the corpus, i.e. COMMAND. This gives a baseline of 57.63%. All trees trained during the experiments were automatically pruned.

6.1 Trees trained on lexical utterance features

The first set of experiments used only lexical features of utterances to predict DAs. From each utterance, up to three features were extracted:

- Base form of "main" element
- Base form of object of main element
- Base form of first word in utterance

Due to some utterances being incomplete or ungrammatical, the parser cannot be relied upon to identify a *main* element. In addition, objects will not always be present. To ensure that at least some information about each utterance is available, the base form of the first word is also always extracted. Manual inspection of the utterances suggests that the first word often conveys some useful information about the purpose of the utterance.

There are two different variables in the experiments:

- The amount of dialogue history taken into account. It important to know whether extracting features from previous utterances can help, and if so, how many such utterances should be considered. The dialogue history has been varied from 0 (i.e. only the current utterance) to 3.
- The number of *distinct* attribute values used. As mentioned previously, only the most popular values for an attribute are likely to provide useful evidence towards classification. To try to discover an "optium" size for the set of values, this was varied from 10 to 40 values for each attribute, plus "misc".

The experimental results are shown in Table 1. The best accuracy achieved by a decision tree classifier was 79.68%, which is a considerable improvement over the baseline accuracy of 57.63%.

	Number of attribute values							
	10		20		30		40	
History Size	DT	NB	DT	NB	DT	NB	DT	NB
0	70.51	71.11	71.91	76.26	75.30	78.49	75.89	78.68
1	77.88	80.48	79.28	85.06	77.07	84.46	75.69	83.27
2	78.09	77.69	79.68	80.87	77.29	80.68	75.50	81.07
3	77.89	76.89	79.08	79.28	77.09	79.88	75.50	80.68

Table 1: Accuracy (%) of decision tree (DT) and naive bayes (NB) classifiers using lexical features

However, in almost all cases, the decision tree classifiers are outperformed by the Naive Baysian classifiers, with the best one reaching an accuracy of 85.06%.

For both types of classifier, the best results are achieved when a dialogue history of 1 is used, with 20 distinct values for each attribute. The values of the attributes are shown in Table 2. However, an interesting observation from the classifiers that use no previous context is the very small increase in the classifier accuracies between using 30 and 40 distinct attribute values. This suggests that the vocabulary employed by users in the majority of their utterances is generally quite limited.

Although the accuracies of the classifiers are boosted by taking into account some dialogue history, the results show that there is no advantage in using a history of size greater than 1 (i.e. considering only features of the immediately preceding utterance in addition to the ones of the utterance to be classified). In the case of the decision tree classifiers, the accuracies remain roughly constant whether a dialogue history of 1, 2 or 3 is used. However, in the case of the Naive Baysian classifiers, there seems to be a definite negative effect of using dialogue histories of a size greater than 1.

These results can probably be explained by the general structure of the dialogues, in which user utterances are most commonly preceded by a system utterance. Most system utterances are drawn from a fixed set of utterances, which are generally uniquely identifiable from the features used in these experiments. The majority of these system utterances condition the user to provide a certain type of response, meaning that the identification of the immediately preceding system utterance is a likely to provide substantial evidence towards the accurate classification of user utterances. Looking further back than the preceding utterance in the dialogue is thus unlikely to provide useful information towards classification.

The above observations may be confirmed by a manual inspection of the decision trees trained during the experiments: the root element in all the trees that use previous context is the first word of the immediately preceding utterance. In some cases, this word alone is enough to predict, or at least severely constrain, the most likely DA tag for the user utterance. Another significant observation from the trees is that the object of the *main* element in the utterances is never used, suggesting that this has little significance in predicting dialogue acts.

6.2 Trees trained on non-lexical features

A second set of experiments was performed using only non-lexical features. Although the previous set of experiments suggest that user utterances generally contain a limited vocabulary, this cannot always be guaranteed. A classifier that uses only non-lexical utterance features would obviously be robust to such variation. The results of these experiments are shown in Table 3. As the number of distinct values of non-lexical attributes obtained from the corpus is quite small (all have less than 20 distinct values), it was not considered necessary to vary the number of distinct attribute values used by the classifiers. The following set of attributes were used:

- Number of words in the utterance
- Morphological tag of the main element
- Surface syntactic tag of the main element
- Morphological tag of object of main element
- Surface syntactic tag of object of main element
- DA tag (for previous utterances)

Current utterance			Preceding Utterance		
First word	Main word	Object	First word	Main word	Object
yes	message	message	end	dictate	message
reply	yes	you	dictate	message	what
er	reply	reply	message	like	subject
next	read	it	what	have	quit
end	send	computer	you	reply	reply
i	like	one	reply	send	you
send	open	read	send	be	it
read	delete	send	finish	finish	send
open	listen	mail	be	want	that
ok	let	what	do	ok	come
erm	thank	call	to	thank	try
delete	be	know	hello	about	money
hm	repeat	i	ok	delete	computer
hi	check	program	er	understand	call
listen	have	come	about	quit	know
thank	finished	go	delete	read	number
please	end	end	hm	collect	make
check	hello	check	more	bye	be
new	go	money	thank	go	com
hello	pick	be	i	let	log

Table 2: Lexical attributes values used in best performing classifiers

History Size	DT	NB
0	68.53	62.35
1	83.07	81.87
2	83.46	80.27
3	83.06	79.28

Table 3: Accuracy (%) of decision tree (DT) and naive bayes (NB) classifiers using non-lexical features

Thee results are encouraging, considering that they are obtained completely independently of lexical information in the utterances. The accuracy of the best classifier (i.e. 83.46%) is only slightly lower than the best lexical-only classifier. The best results in these experiments were achieved by the decision tree classifiers, although only by a relatively small margin. The general pattern of how the accuracy of the classifiers change as the dialogue history increases is very similar to the lexical-only classifiers, i.e. that there is no obvious reason to use a dialogue history of size greater than 1.

A probable reason for the large leap in the accuracy of the classifiers between using no context and using a dialogue history of 1 is the fact that in the latter classifier, the DA tag of the immediately preceding utterance is available. As mentioned previously, the majority of utterances which immediately precede user utterances are system utterances, and the system will "know" the DA of each utterance it produces. In general, all system utterance that are assigned a particular DA tag will condition the same type of response by the user. Therefore, knowing the DA of the immediately preceding utterance should provide a more general and reliable way identifying its characteristics than using only lexical information. Once again, an examination of the decision trees trained during these experiments strengthens this hypothesis: all trees that take context into account have as their root the DA tag of the immediately preceding utterance.

An examination of the best performing decision tree reveals some other interesting features of user utterances. Many nodes further down the tree are concerned with either the length of the current utterance, or the morphological tags assigned to its main element. This suggests that the realisations of different DAs by users have distinctive and consistent non-lexical features. Additionally, in common with the lexical-only classifiers, features relating to the object of the *main* element were never used in the trees trained in this set of experiments.

6.3 Trees trained on lexical and non-lexical features

The results of the previous 2 sets of experiments have shown that classifiers trained on non-lexical features can achieve a level of accuracy that is only slightly lower than those trained on lexical features. It may therefore be expected that combining lexical and non-lexical features will provide a further boost in classifier performance, especially as most DA classifiers described in the literature use a combination of these features. To test this hypothesis, a third set of experiments was carried out. As the decision trees trained in the previous 2 sets of experiments seem to suggest that features relating to the object of the main verb are not relevant in utterance classification, these features were omitted. This leaves the following set of features:

- Number of words in the utterance
- Base form of first word in the utterance
- Base form of "main" word in the utterance
- Mophological tags of main word
- Surface syntax tag of main word
- DA tag (for previous utterances)

For the lexical features (i.e."first word" and "main word"), the attribute set consists of the 20 most popular values from the training corpus. According to the results from the first set of experiments, this seems like an optimal value.

The results obtained are shown in Table 4. The best accuracy was achieved using a Naive Baysian classifier. In common with the results of the previous 2 sets of experiments, this best performing classifier uses a dialogue history of 1. Although the Naive Baysian classifiers trained in this set of experiments are slightly more accurate than their non-lexical counterparts, the results obtained are almost identical to those achieved by the lexical-only classifiers. This seems to suggest that combining lexical and non-lexical features has no particular advantage.

A similar pattern is observable for the decision trees: the results of training trees on a combination of lexical and non-lexical features are consistently slightly better than those trained on lexical features alone, but they are no better than the trees trained on non-lexical features only. This is a desirable outcome as regards the robustness of the classifier, as it suggests that high levels of classification accuracy may be achieved without considering lexical information at all.

History Size	DT	NB
0	75.10	75.50
1	81.87	84.46
2	82.07	80.87
3	81.87	79.28

Table 4: Accuracy (%) of decision tree (DT) and naive bayes (NB) classifiers using lexical and non-lexical features

7 Conclusion

This paper has presented preliminary experimental results of classifiers trained to predict DAs of user utterances in spoken email dialogues. Our domain is much restricted than those used for other classifiers in the literature, meaning that our results are not directly comparable. Our results show that the 2 types of classifiers used, i.e. decision trees and Naive Baysian classifiers, produce roughly comparable results. The former type of classifier seems to do slightly better when only non-lexical information is used, whilst the latter type appears to perform best when lexical information is available. These results must of course be verified using a much larger corpus, which we hope to start collecting soon.

A significant result from the experiments is that the best performing classifier that uses only non-lexical features performs almost as well as the best classifier that uses lexical utterance information. Using such a classifier in the "real" spoken email system would be advantageous for two reasons. Firstly, as mentioned previously, it would be robust to user utterances whose vocabulary varied from the utterances on which the classifier was trained. Such variations could cause a classifier that made use of lexical utterance information to perform badly. The second reason concerns automatic speech recognition (ASR). No matter how good the recognition grammar is, recognition errors will always occur, which will obviously

prove problematic for classifiers that make use of lexical information. However, in the case of word substitution errors, it could be that the recognised word has the same grammatical properties as the word spoken. Thus, if a purely non-lexical classifier were used, such errors made by the ASR would be transparent to it. Of course, a great deal of experimentation with the ASR will be necessary to confirm the validity of this hypothesis.

8 Discussion

Whilst rational dialogue agency theory holds that dialogue behaviour is engendered by intentions and beliefs, and that speech acts, and by extension dialogue acts, are epiphenomenal, they do seem to be a construct that is useful in practical dialogue systems. Speech or dialogue acts are considered as operators that conventionally realize particular kinds of intentions, and the reason sometimes advanced for dispensing with them is the pervasive use of non-conventional implicatures for getting intentions across. However, the possibility of decoding dialogue acts relatively accurately and mostly from surface features and limited dialogue history does suggest that conventional encodings of intentions predominate, at least in our application domain.

Acknowledgements

The work reported here has been co-funded by the European Union and the project partners, in Project No IST-2000-29452: Dynamic Universal Mobility for Adaptive Speech Interfaces. The authors are also grateful to the reviewers for helpful suggestions, in particular the use of cross-validation, and the consideration of two distinct classifiers.

References

H. Bunt and W. Black, editors. 2000. *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics*. John Benjamins.

J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson. 1996. Hcrc dialogue structure coding manual. Technical report, Human Communication Research Centre, University of Edinburgh.

N. Dahlbäck, A. Jönsson, and L. Ahrenberg. 1992. Wizard of oz studies – why and how. In *Third Conference on Applied Natural Language Processing*, Trento, Italy.

S. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast, and J. Quantz. 1995. Dialogue acts in verbmobil. Technical Report 65, Universität Hamburg, DFKI Saarbrücken, Universität Erlangen, TU Berlin.

T. Mitchell. 1997. *Machine Learning*. McGraw-Hill.

N. Reithinger and M. Klesen. 1997. Dialogue act classification using language models. In *Proceedings of Eurospeech '97*, pages 2235–2238, Rhodes, Greece.

K. Samuel, S. Carberry, and K. Vijay-shanker. 1998. Computing dialogue acts from features with transformation-based learning. In *Proceedings of the AAAI 98*, pages 90–97.

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. *Computational Linguistics 26(3)*.

I. Witten and E. Frank. 2000. *Data mining: Practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, San Francisco.