# CROSS-DOMAIN CLASSIFICATION USING GENERALIZED DOMAIN ACTS

*Andrew N. Pargellis and Alexandros Potamianos*

Bell Labs, Lucent Technologies, 600 Mountain Ave., Murray Hill, NJ 07974, U.S.A.
email: {anp,potam}@research.bell-labs.com

## ABSTRACT

**Cross-domain classification for speech understanding is an interesting research problem because of the need for portable solutions in the design for spoken dialogue systems. In this paper, a two-tier classifier is proposed for speech understanding. The first tier consists of domain independent dialogue acts while the second tier consists of application actions that are domain specific. A maximum likelihood and a minimum classification error formulation are proposed for the first tier of the classifier, i.e., for dialogue act classification. The performance of the classifier is investigated for three application domains. Cross-domain classification error is two to four times higher than in-domain classification error. A 10-15% reduction in cross-domain classification error rate is achieved by adding generic domain independent training data for each dialogue act and by mapping words to semantic concepts.**

## 1. INTRODUCTION

Much of the development time of spoken dialogue systems is spent on data collection, annotation, and on authoring understanding models for each new application domain. Dialogue acts are an attractive domain independent semantic representation of the user's input. Due to their generality, dialogue acts are often used in dialogue management [7, 8]. However, dialogue acts are not equally popular for speech understanding; application dependent actions, rather than dialogue acts, are typically used for understanding. Work on dialogue act classification can be found in the literature [15, 12, 1]. Little work exists, however, on investigating the portability of dialogue act models across application domains. Although some dialogue act definitions are application independent, their usefulness for speech understanding will be limited if statistical models have to be trained anew for each domain.

Our goal in this paper, is to improve classification accuracy and portability of understanding models across domains, and to speed up the process of building understanding models for new applications. A two step classification algorithm is proposed: user utterances are first classified into dialogue acts and then into application dependent actions. In this paper, we focus on dialogue act classification across different domains and investigate the portability of the dialogue act models. In the next section a two-tier classifier architecture is presented. In Section 3, maximum likelihood and minimum classification error training are proposed for dialogue act classification. Dialogue acts are defined in Section 4. Finally, model portability is investigated for three different domains and ways for improving cross-domain performance are proposed.

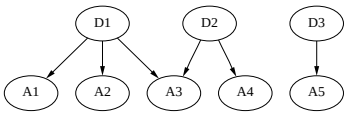


Figure 1: An example of the two-tier classifier: 'D' denotes dialogue acts and 'A' denotes application actions.

## 2. CLASSIFIER ARCHITECTURE

The proposed classifier consists of two tiers as shown in Fig. 1. The first tier consists of dialogue acts D1, D2 ... which are common across all application domains, while the second tier consists of application actions A1, A2, ... that are domain specific. Note that an application action might correspond to more than one dialogue act, e.g., A3 in Fig 1. The classification of a user utterance to one (or more) application action(s) happens in two stages; first the utterance is classified to a dialogue act and then to an application action. Specifically, using the maximum likelihood decoding formulation

$$\max_A P(A|s) = \max_A \sum_D P(A, D|s)$$
$$= \max_A \sum_D P(D|s)P(A|D, s) \quad (1)$$

where $A$ is a application action, $D$ is a dialogue act, and $s$ is the user's utterance.

In this paper, models are proposed for computing $P(D|s)$. The portability of these models across domains is evaluated and ways to improve cross-domain classification performance are proposed. Although models for computing the second term $P(A|D, s)$ are not investigated in this paper, such models should be simple and require little training data.

## 3. CLASSIFICATION ALGORITHMS

As discussed in the previous section the posterior probability $P(D_k|s)$ has to be computed for each dialogue act $D_k$. A typical statistical approach to this problem involves constructing a model $L_k$ for each dialogue act $D_k$ from the training set $I_k$ using a maximum likelihood learning criterion and then determining the dialogue act from the user input $s$ as:

$$\hat{k} = \operatorname{argmax}_k P(L_k|s) = \operatorname{argmax}_k \{P(s|L_k)P(L_k)\}. \quad (2)$$

If the user input is given as a text string then $I_k$ is a set of transcribed sentences that belong to dialogue act $D_k$. A simple statistical model for $I_k$ is the computation of the word sequence probability corresponding to the user's utterance. For this purpose we have used the Variable Ngram Stochastic Automaton [13]. If $L_k$ is the $n$-gram statistical model trained from $I_k$ and the input utterance $s = w_1 w_2 .. w_N$ is represented as $\bigoplus_n w_n$ then

$$P(L_k|s) \approx P(L_k| \bigoplus_{n:w_n \in L_k} w_n) \, [(c_{oov})^{\sum_n \delta(w_n \notin L_k)}] \qquad (3)$$

where $w_n \in L_k$ signifies that word $w_n$ is in vocabulary drawn from $L_k$, $\delta(w_n \notin L_k) = 1$ for out of vocabulary (OOV) word (else 0) and $c_{oov}$ is a task dependent constant penalty for deletion of OOV words from input $s$. The selected dialogue act $D_{\hat{k}}$ is the one that maximizes the probability given in Eq.(3). The existence of OOV words in the transcribed input string $s$ is common for closed vocabulary systems. Moreover, OOV words might appear even when $s$ is the output of an automatic speech recognizer because in general the training corpus $\mathcal{I}_k$ for understanding model $L_k$ is a subset of the language model training corpus. A more detailed discussion of the understanding model can be found in [10].

### 3.1. Class-based Classifier

Before training $n$-gram models words and phrases in the utterances were first mapped using 23 semantic classes (which encompass three domains: movie, travel, and computer game): Airlines, Airports, Alphabet, Cars, CityNames, Colors, CreditCards, FirstPerson, FillerWords, FirstName, Hotels, Months, MovieTitles, Numbers, Objects, Organizations, Region, SecondPerson, StateNames, TheatreNames, ThirdPerson, Times, and Weekdays. The semantic parsing improved results significantly as discussed in Section 6, because more accurate statistics can be computed for a class (e.g. 'Movie') than specific instances of the class (e.g. a specific movie title).

### 3.2. Discriminative Training

Certain features are much more important than others in the dialogue act classification process. For example 'where' and 'when' are important cues for 'Req_Location' and 'Req_Time' dialogue acts respectively, while filler words like 'the' and 'a' are not very useful features. Maximum likelihood training often performs poorly on sparse data and is not able to capture the discriminative power of features. To improve classification, *class-independent* exponential weights $\gamma$ are introduced in the statistical $n$-gram model. Specifically, assuming a bigram model,

$$P(s|L_k) \approx \prod_{n=1}^{N} P(w_n|w_{n-1}, L_k)^{\gamma(w_{n-1}, w_n)} \qquad (4)$$

Note that the weights are a function of the current word and word history, but independent of the class $k$. Weights are trained via gradient descent to maximize a class separation measure (e.g. [4]):

$$P(s|L_k) - \frac{1}{N-1} \sum_{i \neq k} P(s|L_i), \qquad (5)$$

where $k$ is the correct understanding class (out of a total of $N$ classes).

## 4. DIALOGUE ACT DEFINITIONS

Dialogue and speech acts, as traditionally defined in the literature, capture the semantic and pragmatic content of an utterance[5]. However, the state of the art in dialogue act classification is based on word $n$-grams, a statistical model that captures mostly lexical and (some) syntactic information. As a result, classification performance is very poor for dialogue acts that have high lexical variability in their realizations, e.g., "clarify," "digress," "motivate."

Alternatively, dialogue acts can be defined to both capture semantic/pragmatic content and minimize intra-act lexical variability. In this paper, we adopt this approach when defining dialogue acts. The set chosen is similar to that used by the VERBMOBIL project[12, 7] and other groups[8, 6], but some ambiguous dialogue acts were discarded. The set of 12 dialogue acts selected is shown below. Examples are shown in parentheses after the description.

**Accept**: user accepts System's suggestion ("Yes")
**Bye**: terminate the present subdialogue ("I'm done.")
**Greet**: ("Hello")
**Init**: user initiates a new dialogue ("I want to make a plane reservation")
**Reject**: user rejects System's suggestion ("No")
**Req_Action**: command System to do some action ("Put the clue in my bag")
**Req_Location**: ("Where is Pulp Fiction playing?")
**Req_Suggest**: general question that is not a Req_Loc or Req_Time ("Is there a United flight in the morning?")
**Req_Time**: ("When does the next flight leave?")
**Req_YN**: general question that takes a Yes or No answer.
**Suggest**: user answers a System question, ("3 p.m.")
**Thank**: user acknowledges a System action

There were three additional dialogue acts whose utterances were not included in the understanding tests. Depending on the domain, these three dialogue acts together accounted for 4 to 12% of the total utterances.

**Garbage**: unintelligible or unimportant discourse, fragments, or back-channeling ("I see," "uh uh," "can I go to"). About 4%.
**Multi_Tags**: multiple dialogue acts in a single utterance ("Yes, I'll take the first flight, and I also will need a hotel reservation," "Thank you. Goodbye."). Ranged from 0 to 8%.
**Unknown**: ambiguous utterances that could not be identified out of context. For example, "Then I can drive to Miramar" could be either a question (Req_YN) or a 'Suggest' or even 'Garbage.' A few tenths of a percent were in this category.

## 5. DESCRIPTION OF EXPERIMENTS

Three domains were selected for experimentation. The three domains referred to as 'Carmen', 'Movie', and 'Travel' contained a different mix of dialogue acts and lexical content. 'Carmen' (short for 'Where in the USA is Carmen-Sandiego?') is data collected in a Wizard-of-oZ (WoZ) experiment where children used voice to play this computer game [11]. 'Carmen' has a limited set of commands and therefore a somewhat constrained dialogue. 'Movie' is data collected from a spoken dialogue movie information system [2]. 'Movie' has a limited set of dialogue acts, but the dialogue was more open-ended than 'Carmen.' Finally, 'Travel' is data collected in a WoZ experiment for a travel reservation spoken dialogue system [14]. This domain contained the most open-ended dialogue and proved to be most challenging for dialogue act classification.

| | Carmen | Movie | Travel |
|---|---|---|---|
| Accept | 1.5 % | 1.4 % | 26.1 % |
| Bye | 1.6 | 1.1 | 2.2 |
| Greet | 0.8 | 0.7 | 1.2 |
| Init | 29.1 | 1.2 | 6.1 |
| Reject | 0.9 | 0.7 | 5.8 |
| Req_Action | 18.1 | 10.7 | 7.5 |
| Req_Location | 17.5 | 22.5 | 3.2 |
| Req_Suggest | 20.4 | 31.1 | 9.6 |
| Req_Time | 0.7 | 22.7 | 2.1 |
| Req_YesNo | 1.7 | 2.6 | 4.7 |
| Suggest | 3.2 | 4.1 | 28.3 |
| Thank | 4.0 | 0.7 | 2.4 |

Table 1: The distribution, in percent, of the 12 dialogue acts (rows) for each of the three domains (columns) studied.

### 5.1. Testing and Training Sets

The total number of utterances in each domain were: 'Carmen' (2416), 'Movie' (2500), and 'Travel' (1593). These numbers include utterances from the three dialogue acts later filtered out: Garbage, Multi-Tags, and Unknown. The data were divided equally in half for training and testing, respectively.

N-gram models were trained for each of the 12 dialogue acts [10]. A bigram understanding model was used and the out-of-vocabulary penalty in Eq.(3) was set to 4 for both in-domain and cross-domain experiments (lower OOV penalty gave somewhat better results for mismatched training and testing conditions).

It is necessary to have at least one training utterance per dialogue act and domain to be able to build a dialogue act model. When a dialogue act was missing from a domain, e.g., 'Accept' for the 'Movie' domain, a small set of domain independent generic utterances was generated for that dialogue act. For example, for the 'Accept' dialogue act we used five instances of each of the words in the set: 'fine', 'okay', 'sure', 'yes', for a total of 20 'Accept' utterances. This set was appended to the training set of that domain.

An important question is the amount of overlap of dialogue acts among applications. In Table 1 the distribution of the 12 dialogue acts is shown for each of the three domains (columns). The table shows some major differences between the three domains. For example, the 'Travel' domain contains over half of the total number of utterances classified in the 'Accept' and 'Suggest' dialogue acts; both dialogue acts were rarely used in the 'Carmen' and 'Movie' domains. Overlap was greater between the 'Carmen' and 'Movie' domains, although 'Carmen' had 29% 'Init' and 'Movie' had 23% 'Req_Time', both of which were poorly represented in the other domains.

## 6. RESULTS AND DISCUSSION

The understanding accuracies for the three different domains based on the bigram understanding model are summarized in Table 2. Note that chance is 8.3 % (one out of twelve). The seven combinations of training conditions are listed in the left-hand column. The three testing sets correspond to the three tasks which are shown in the columns. Only the understanding accuracies for the bigram understanding model are shown. The accuracies for a trigram model were essentially the same as those for the bigram, while the unigram accuracies were usually a few percentage points lower.

| | Carmen | Movie | Travel |
|---|---|---|---|
| Carmen | 91.3 % | 76.6 % | 44.8 % |
| Movie | 71.6 | 96.9 | 45.3 |
| Travel | 60.2 | 86.6 | 78.0 |
| Carmen-Movie | 91.7 | 97.4 | 45.6 |
| Carmen-Travel | 89.6 | 74.9 | 80.3 |
| Movie-Travel | 63.1 | 96.7 | 80.4 |
| All | 89.9 | 97 | 84.3 |

Table 2: The bigram understanding accuracy, in percent, for each of the three domains tested (columns). Seven combinations were used for the training sets (rows). 'All' includes utterances from all three domains for training.

In-domain results (train.test = carmen.carmen, etc) are above 90% except for the 'Travel' domain. The 'Movie' domain was the easiest to recognize using data from any combination of domains for training. This is probably because of the good coverage of the dialogue acts predominantly used in 'Movie' (see Table 1). Utterances for the 'Travel' domain were the most difficult to classify since the overlap was especially poor between the 'Travel' domain and the others. Note that adding more data from different domains had mixed results.

Of particular interest are the three cross-domain cases, with associated classification scores: carmen-movie.travel (45.6%), carmen-travel.movie (74.9%), and movie-travel.-carmen (63.1%). These indicate the ability of a domain to classify dialogue acts in a totally new domain and therefore enable an application developer to design applications for new domains [9]. However, the difference between in-domain and cross-domain classification accuracy is large: errors rates are typically 2-4 times larger for the mismatched conditions. The poor results for cross-domain classification are due to the mismatch between different domains, both in terms of distribution of dialogue acts (see Table 1) and distribution of words/$n$-grams. Certain dialogue acts are especially hard to classify because of highly domain dependent lexicalization, e.g., 'Suggest.'

To investigate the mismatch between the various domains the Kullback-Leibler (K-L) distance [3] between the understanding bigram models of the three domains was computed for each dialogue act. The largest distance was between the 'Travel' and 'Carmen' domains, closely followed by 'Travel' and 'Movie'; the distance between 'Carmen' and 'Movie' was about three times smaller. The K-L distance can help explain the poor cross-domain classification results in Table 2.

### 6.1. Towards Improving Classification Accuracy

To improve the classification (understanding) accuracy a set of generic dialogue acts were included in the training set for each domain and words were mapped into semantic categories. The results for various testing and training conditions are shown in Fig. 2. The four conditions (Cases) listed on the horizontal axis are as follows: Case 1: a single generic (domain independent) utterance was added to the training set for each dialogue act in each domain, no mapping of words to semantic categories. Case 2: as for Case 1, but with mapping of words into the 23 semantic categories specified in Section 3.1. Case 3: a set of twenty generic utterances for each dialogue act (four utterances repeated five times) added to the training set), no mapping. Case

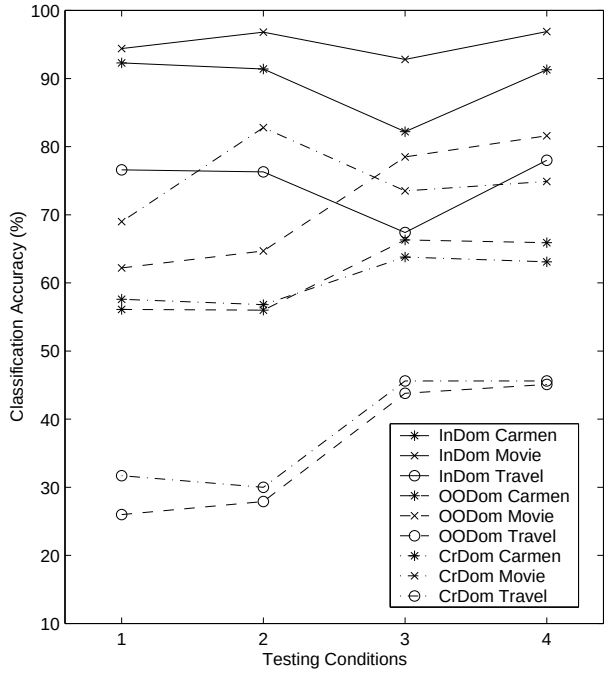


Figure 2: The classification accuracy, in percent, for each of the three domains. Marker types define the testing domain: "carmen" (star), "movie" (x), "travel" (circle). Line types define the training/testing conditions: "in-domain" (solid), "average of one-out-of-domain" (dashed), "combination of two-out-of-domain" (dashed dotted).

4: as for Case 3, but with semantic mapping. Note that the data in Table 2 correspond to Case 4.

Most of the improvement for cross-domain classification was due to the addition of the generic domain independent utterances in the training set for each dialogue act. Mapping words to semantic concepts gave little additional improvement; most of the improvement was for in-domain classification. Overall, adding generic utterances and concept in the training improved cross-domain classification accuracy by 5-15%, while in-domain classification accuracy was pretty much unchanged (compare Cases 1 and 4).

Finally, throwing out common filler words, e.g., 'the', from the training and test data resulted in an additional cross-domain classification improvement of about 3%. A simple implementation of the discriminative training (see Section 3.2) gave no significant additional improvement in classification accuracy. It is interesting to note that, as expected, filler words were judged to be the least useful features by the discriminative training procedure and were least weighted in the statistical model of Eq.(4).

## 7. SUMMARY

A two-tier architecture for application action classification was introduced. Maximum likelihood $n$-gram based classification models were proposed for the first-tier, which classifies utterances into domain independent dialogue acts. In-domain and cross-domain classification accuracy was investigated for a set of twelve dialogue acts for three applications: computer gaming, movie information and travel reservation. The average in-domain classification accuracy was 87%, ranging from 78% to 97%. The classification across domains (mismatched training and testing) was significantly lower, averaging 64%, ranging from 45% to 87%. Combining the training data for all three domains resulted in classification accuracies comparable with or better than the matched conditions. Adding generic domain independent utterances in the training set and mapping words to concepts significantly improved classification accuracy. More research is needed to improve cross-domain classification performance and to investigate adaptation of understanding models across application domains.

**Acknowledgments**: The authors would like to thank Dr. Alex Rudnicky for supplying the CMU Communicator data and Dr. Shrikanth Narayanan at AT&T Labs for providing the 'Carmen Sandiego' data.

## 8. REFERENCES

[1] J. Chu-Carroll, "A statistical model for discourse act recognition in dialogue interactions," in *Working Notes of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pp. 12–17, 1998.

[2] J. Chu-Carroll, "MIMIC: An adaptive mixed initiative spoken dialogue system for information queries," in *Proceedings of the 6th ACL Conference on Applied Natural Language Processing*, Seattle, Washington, May 2000.

[3] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., New York, 1991.

[4] B-H. Juang and S. Katagiri, "Discriminative learning from minimum error classification," in *IEEE Trans. on Signal Proc.*, vol. 40, no. 12, pp. 3043-3054, 1992.

[5] D. Jurafsky, J. H. Martin, *Speech and Language Processing*, Prentice-Hall Inc., Upper Saddle River, 2000.

[6] D. Jurafsky et al., "Automatic Detection of Discourse Structure for Speech Recognition and Understanding," in *Proc. IEEE Workshop on Speech Recognition and Understanding*, Santa Barbara, California, 1997.

[7] E. Maier, "Context Construction as Subtask of Dialogue Processing - the VERBMOBIL Case," TWLT 11, Twente, Netherlands, June 1996.

[8] D. G. Novick, S. Sutton, "Building on Experience: Managing Spoken Interaction Through Library Subdialogues," TWLT 11, Twente, Netherlands, June 1996.

[9] A. Pargellis, J. Kuo, C.-H. Lee, "Automatic Dialogue Generator Creates User Defined Applications," in *Proc. European Conf. on Speech Communication and Technology*, Budapest, Hungary, Sept. 1999.

[10] A. Potamianos, G. Riccardi, S. Narayanan, "Categorical Understanding Using Statistical Ngram Models," *Proc. European Conf. on Speech Communication and Technology*, Budapest, Hungary, Sept. 1999.

[11] A. Potamianos and S. Narayanan, "Spoken dialog Systems for Children," in *Proc. Intl. Conf. on Acoustic Speech and Signal Processing*, pp. 197-201, Seattle, Washington, 1998.

[12] N. Reithinger, M. Klesen, "Dialogue Act Classification Using Language Models," *Proc. European Conf. on Speech Communication and Technology*, Rhodes, Greece, Sept. 1997.

[13] G. Riccardi, R. Pieraccini and E. Bocchieri, "Stochastic Automata for Language Modeling," Computer Speech and Language, vol. 10(4), pp. 265-293, 1996.

[14] A. Rudnicky and W. Xu, "An agenda-based dialog management architecture for spoken language systems," in *Proc. Workshop on Automatic Speech Recognition and Understanding*, Keystone, Colorado, Dec. 1999.

[15] A. Stolcke et al, "Dialog Act Modeling for Conversational Speech," in *Proc. AAAI Spring Symposium*, Stanford, California, Mar. 1998.