MODEL ADAPTATION FOR DIALOG ACT TAGGING

Gokhan Tur *Umit Guz*[*] *Dilek Hakkani-Tür*

SRI International
Menlo Park, CA 94025
{gokhan,umit}@speech.sri.com

International Computer Science Institute (ICSI)
Berkeley, CA 94704
dilek@icsi.berkeley.edu

ABSTRACT

In this paper, we analyze the effect of model adaptation for dialog act tagging. The goal of adaptation is to improve the performance of the tagger using out-of-domain data or models. Dialog act tagging aims to provide a basis for further discourse analysis and understanding in conversational speech. In this study we used the ICSI meeting corpus with high-level meeting recognition dialog act (MRDA) tags, that is, *question*, *statement*, *backchannel*, *disruptions*, and *floor grabbers/holders*. We performed controlled adaptation experiments using the Switchboard (SWBD) corpus with SWBD-DAMSL tags as the out-of-domain corpus. Our results indicate that we can achieve significantly better dialog act tagging by automatically selecting a subset of the Switchboard corpus and combining the confidences obtained by both in-domain and out-of-domain models via logistic regression, especially when the in-domain data is limited.

1. INTRODUCTION

Dialog act tagging is a basic building block for spoken language understanding in human/human conversations or multiparty meetings. A dialog act is an approximate representation of the illocutionary force of an utterance, such as question or backchannel [1]. Dialog acts are designed to be task independent by definition. The main goal of dialog acts is to provide a basis for further discourse analysis and understanding. For example, dialog acts can be used to extract the question/answer pairs in a meeting. Note that dialog acts can be organized in a hierarchical fashion. For instance statements can be further categorized as *command* or *suggestion*. There are a number of predefined dialog act sets in the literature, such as DAMSL [2] and MRDA [3].

In this study we used the ICSI meeting corpus with high-level MRDA tags, i.e. *question*, *statement*, *backchannel*, *disruptions*, and *floor grabbers/holders* [3]. Backchannels are short phrases such as *yeah* or *uh huh* to indicate that the listener is actually following the speaker. Floor grabbers indicate that the person wants to start talking; similarly floor holders indicate that the speaker is not yet finished. Disruptions include the statements uncompleted for some reason. Below is an example dialog along with dialog acts:

- Speaker 1: So is this OK with you? (**question**)
- Speaker 2: Yes, (**statement**) but I do- (**disruption**)
- Speaker 1: Come on (**floor grabber**) I want this very much (**statement**)
- Speaker 2: Uh huh (**backchannel**)
- Speaker 1: And I want ... (**statement**)

Dialog act tagging is generally framed as an utterance classification problem [1, 4]. Large amounts of in-domain data are usually transcribed, segmented into dialog acts, and then labeled manually, an expensive and laborious process. The problem is that although dialog acts are designed to be task independent and even though we consider only five top level dialog acts, there are still significant differences between different corpora due to different dialog act distributions and labeling inconsistencies.

In cases where only a limited amount of dialog act annotated data is available, an immediate solution would be to use adaptation methods with existing out-of-domain dialog act data or models. Although statistical model adaptation has been a well studied area in speech recognition for acoustic and language modeling [5, 6, 7], there is comparably less work done on natural language processing. One recent study is on the adaptation of natural language understanding using a common adaptation method of *maximum a posteriori* (MAP) adaptation [8], which adapts the hidden vector state model built for ATIS application to DARPA Communicator. Another study is about supervised and unsupervised adaptation of probabilistic context-free grammars to a new domain using again MAP adaptation [9]. In our previous study, we proposed model adaptation methods for call classification in a goal-oriented spoken dialog system used for customer care [10].

Previously, Venkataraman *et al.* tried employing active learning and lightly supervised learning for dialog act tagging. They concluded that while active learning does not help significantly for this task, exploiting unlabeled data

[*] on leave from the Isik University, Istanbul, Turkey

by using minimal supervision is effective in certain conditions [11, 12]. Note that in this study we consider only supervised model adaptation and analyze the effect of adaptation in a controlled setting where the in-domain data is from the ICSI meeting corpus, and out-of-domain data is the Switchboard (SWBD) corpus with the SWBD-DAMSL tag set [13]. A research challenge with this out-of-domain data is that the floor grabbers/holders are not considered to be a separate class; hence, there are only four top-level dialog acts.

In the following section, we briefly explain our approach. Then, in Section 3, we present the experiments and results.

2. APPROACH

The aim of supervised adaptation is to exploit the existing labeled data and models from previous corpora or applications for improving the performance of the new similar applications, which generally have a lesser amount of labeled data. The idea is adapting the existing model using the smaller amount of already-labeled data from the new application, thus reducing the amount of human-labeling effort necessary to come up with decent statistical systems.

The simplest way of exploiting the existing labeled data from a similar application is data concatenation, where the new model is trained using the data from the previous application concatenated to the data labeled for the new application. For example, for the language modeling task this is actually equivalent to count mixing using equal weights. Count mixing is also shown to be equivalent to model interpolation, which is equivalent to MAP adaptation [14, 15]. During model interpolation, an out-of-domain model, θ_{OOD}, is interpolated with an in-domain model, θ_{ID}, to form an adapted model, $\hat{\theta}$:

$$P_{\hat{\theta}}(w_i|h_i) = \gamma \times P_{\theta_{OOD}}(w_i|h_i) + (1-\gamma) \times P_{\theta_{ID}}(w_i|h_i) \quad (1)$$

where $P_\theta(w_i|h_i)$ is the probability of the current word w_i given the history of $n-1$ words, h_i, in an n-gram language model θ. γ is the weight usually estimated using a development set. There are a number of ways to do this estimation. While one can simply try out all the possible values using a development set, another option would be training linear or logistic regression models with the development set.

All these methods and definitions actually hold for statistical classification models. One can apply Equation 1 by simply using the scores or confidences obtained by the classification model. In this study we use the Boosting family of classifiers, which are shown to be very effective for text classification. Boosting is an iterative procedure; on each iteration a weak classifier is trained on a weighted training set, and at the end, the weak classifiers are combined into a single classifier. Each weak classifier (e.g. "decision stump") checks the absence or presence of a feature. In our study, we use only lexical information as features, that is, word n-grams. Note that our approach is independent of the specific classification algorithm used.

While computing the interpolation weights, one check out all possible weights manually, but a more principled approach would be training regression models for this purpose. Linear regression will directly provide the weights for each of the models. We also tried training logistic regression models for each of the five high-level dialog acts using a development set with the Newton-Raphson Method [16]. Then the interpolation formula can be represented by the logistic function:

$$C_{\hat{\theta}}(W) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \times C_{\theta OOD}(W) + \beta_2 \times C_{\theta ID}(W)}}$$

where the β values are the regression weights, and $C_M(W)$ is the confidence given to the utterance W by the model M.

Another method we propose is using the scores or confidences obtained by the out-of-domain model as an additional feature during training. This is a straightforward method especially for discriminative classification methods such as Boosting or SVM, which can exploit continuous valued features effectively.

One problem with our case is that the dialog act classes are not the same. The Switchboard corpus does not have any utterances labeled as floor grabbers/holders. There are a number of ways to cope with this problem. We chose the method of "exclusion": we extracted all the utterances with this tag from a 5,000-utterance portion of the ICSI corpus. This is the minimum amount of in-domain labeled data we assume in the experiments. Then we excluded these utterances from the SWBD corpus. This reduced the amount of SWBD data by 19% due to some floor grabbers/holders which frequently appear in the data, such as *yeah*, not necessarily with this tag exclusively.

3. EXPERIMENTS AND RESULTS

We performed controlled experiments for analyzing the effectiveness of the dialog act model adaptation. We drew learning curves by changing the size of the available in-domain (ICSI) data while keeping the out-of-domain data (SWBD) constant.The data properties are shown in Table 1. In addition to the test set we reserved a portion of the ICSI corpus for tuning the interpolation weights. More specifically, we used 51 meetings for training, 11 meetings for tuning, and 11 meetings for testing as in [17]. As seen from the table, ICSI meeting utterances are much shorter, maybe because of visual contact between the speakers, or because there are typically more than two speakers in a meeting unlike the case of the telephone conversations of the Switchboard corpus. All the experiments are done using

	ICSI	SWBD
Training Data Size	80577 utt.	64874 utt.
Test Data Size	16211 utt.	N/A
Dev Data Size	16501 utt.	N/A
Average Utterance Length	7.58	10.45
Questions	6%	4%
Disruptions	12%	6%
Floor Grabbers/Holders	10%	0%
Statements	58%	58%
Backchannels	12%	29%

Table 1. Data characteristics used in the experiments.

Training Set	Test Set	
	ICSI	SWBD
ICSI	22.02%	30.17%
SWBD	42.63%	N/A
ICSI+SWBD	24.58%	N/A
ICSI5K	25.87%	31.59%
ICSI5K+SWBD	31.24%	N/A

Table 2. Baseline results for the experiments. ICSI5K is a 5,000-utterance subset of the ICSI corpus

manually trancribed and segmented data in order not to deal with automatic speech recognition and sentence segmentation noise. We performed our tests using the Boostexter tool [18]. For all experiments, we used word trigrams as features and iterated 1000 times. We did not use any contextual information, such as the previous or the following dialog act tag, which may improve the performance further.

In this experiment, the goal is adapting the classification model for ICSI data using SWBD so that the resulting model for ICSI would perform better. Table 2 presents the baseline results using training and test data combinations. The rows indicate the training sets, and columns indicate the test sets. The values are the classification error rates, which are the ratios of the utterances for which the classifier's top scoring class is not one of the correct intents. As seen, although the two corpora are very similar, when the training set does not match the test set, performance drops drastically. The third row is simply the concatenation of both training sets (indicated by "+"). Adding SWBD training data to ICSI does not help; actually, it hurts significantly. Since we expect the proposed adaptation method to work better with less application specific training data, we also report results using 5,000 examples from ICSI, called ICSI5K. With this small amount, actually, concetenation hurt more since now out-of-domain data became dominant.

Then we tried adaptation using only ICSI5K as shown in Table 3. We tried linear and logistic regression methods as well as the method in which we use confidences obtained

Adaptation	Error Rate
ICSI5K (Feature)	25.27%
ICSI5K (Log-Regress)	24.81%
ICSI5K (Lin-Regress)	25.39%

Table 3. Adaptation results with various methods. "Feature" indicates using confidences as an additional feature during training. "Log-Regress" and "Lin-Regress" indicate learning the weights via logistic and linear regression, respectively. "+" indicates simple concatenation of data sets.

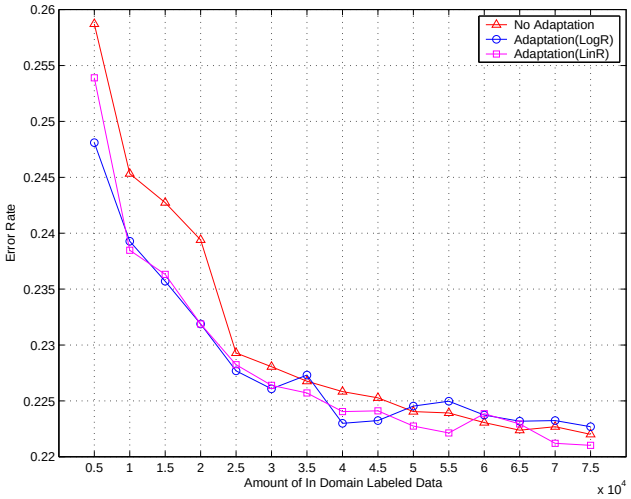

Fig. 1. Results using dialog act classification model adaptation. The top learning curve is obtained using just ICSI MRDA data as a baseline. The lower learning curves are obtained using the adaptation with the Switchboard corpus where the weights are trained using linear and logistic regression.

by the SWBD model as additional feature while training. Using logistic loss performed the best for this experiment. Note that an improvement of around 0.6% is significant according to the Z-test for a 95% confidence interval.

We have also drawn the learning curves as presented in Figure 1. The top-most curve is obtained using random selection of only ICSI training data. The lower curves are obtained using linear and logistic regression. Using out-of-domain confidences as features did not help for data sizes more than 5,000 in-domain utterances. When we employ adaptation with only 5,000 utterances from ICSI, we have seen more than 1% absolute improvement. This improvement reduces, as expected, as we increase the amount of in-domain data. But still for 10,000 utterances we achieve the same performance obtained by 20,000 in-domain utterances, a factor of 2 reduction in the amount of data needed. We can improve the performance significantly by exploiting the Switchboard data when the in-domain data size is

less than 25,000. After about 50,000 in-domain utterances the gain disappears completely, as expected.

4. CONCLUSIONS AND DISCUSSION

We have presented a supervised adaptation method for dialog act tagging. We have shown that, for this task, it is possible to boost the performance of the tagger when there is not much training data available. Our results indicate that we have achieved the same classification accuracy using around 50% less labeled data.

It is also possible to apply the same idea for other data-driven speech and language processing tasks that may need adaptation such as topic classification, named entity extraction, or sentence segmentation.

Our future work includes unsupervised adaptation of dialog act classification models. This will enable us to bootstrap new dialog act models without labeling any application-specific data. Another venue is combining dialog act tagging with dialog act segmentation and employing adaptation in a combined fashion.

Acknowledgments: This material is based upon work supported by the Scientific and Technological Research Council of Turkey (TUBITAK) and Defense Advanced Research Projects Agency (DARPA) GALE (HR0011-06-C-0023) and CALO (NBCHD-030010) funding at ICSI and SRI, respectively. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA. We thank Andreas Stolcke and Elizabeth Shriberg for many helpful discussions.

5. REFERENCES

[1] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[2] M. Core and J. Allen, "Coding dialogs with the DAMSL annotation scheme," in *Proceedings of the Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, Cambridge, MA, November 1997.

[3] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI Meeting Recorder Dialog Act (MRDA) Corpus," in *Proceedings of the SigDial Workshop*, Boston, MA, May 2004.

[4] M. Zimmermann, D. Hakkani-Tür, E. Shriberg, and A. Stolcke, "Text based dialog act classification for multiparty meetings," in *Proceedings of the MLMI*, Washington D.C., May 2006.

[5] G. Riccardi and A. L. Gorin, "Stochastic language adaptation over time and state in a natural spoken dialog system," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, pp. 3–9, 2000.

[6] M. Bacchiani, B. Roark, and M. Saraclar, "Language model adaptation with MAP estimation and the perceptron algorithm," in *Proceedings of the HLT-NAACL*, Boston, MA, May 2004.

[7] V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of gaussian mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, 1995.

[8] Y. He and S. Young, "Robustness isuues in a data-driven spoken language understanding system," in *Proceedings of the HLT/NAACL Workshop on Spoken Language Understanding*, Boston, MA, May 2004.

[9] B. Roark and M. Bacchiani, "Supervised and unsupervised PCFG adaptation to novel domains," in *Proceedings of the HLT-NAACL*, Edmonton, Canada, May 2003.

[10] G. Tur, "Model adaptation for spoken language understanding," in *Proceedings of the ICASSP*, Philadelphia, PA, May 2005.

[11] A. Venkataraman, A. Stolcke, and E. E. Shriberg, "Automatic dialog act tagging with minimal supervision," in *Proceedings of the Australian International Conference on Speech Science and Technology*, Melbourne, Australia, December 2002.

[12] A. Venkataraman, Y. Liu, E. Shriberg, and A. Stolcke, "Does active learning help automatic dialog act tagging in meeting data?," in *Proceedings of the Interspeech*, Lisbon, Portugal, September 2005.

[13] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard SWBD-DAMSL labeling project coder's manual," Tech. Rep. 97-02, University of Colorado Institute of Cognitive Science, 1997.

[14] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *Proceedings of the ICSLP*, Denver, CO, September 2002.

[15] M. Bacchiani and B. Roark, "Unsupervised language model adaptation," in *Proceedings of the ICASSP*, Hong Kong, April 2003.

[16] A. Agresti, *Categorical Data Analysis*, chapter 4, pp. 84–117, John Wiley and Sons, 1990.

[17] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proceedings of the ICASSP*, Philadelphia, PA, March 2005.

[18] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2/3, pp. 135–168, 2000.