Dialogue Act Recognition using Reweighted Speaker Adaptation

Congkai Sun
Institute for Creative Technologies
12015 Waterfront Drive
Playa Vista, CA 90094-2536
csun@ict.usc.edu

Louis-Philippe Morency
Institute for Creative Technologies
12015 Waterfront Drive
Playa Vista, CA 90094-2536
morency@ict.usc.edu

Abstract

In this work we study the effectiveness of speaker adaptation for dialogue act recognition in multiparty meetings. First, we analyze idiosyncracy in dialogue verbal acts by qualitatively studying the differences and conflicts among speakers and by quantitively comparing speaker-specific models. Based on these observations, we propose a new approach for dialogue act recognition based on reweighted domain adaptation which effectively balance the influence of speaker specific and other speakers' data. Our experiments on a real-world meeting dataset show that with even only 200 speaker-specific annotated dialogue acts, the performances on dialogue act recognition are significantly improved when compared to several baseline algorithms. To our knowledge, this work is the first [1] to tackle this promising research direction of speaker adaptation for dialogue act recogntion.

1 Introduction

By representing a higher level intention of utterances during human conversation, dialogue act labels are being used to enrich the information provided by spoken words (Stolcke et al., 2000). Dialogue act recognition is a preliminary step towards deep dialogue understanding. It plays a key role in the design of dialogue systems. Besides, Fernandez et al. (2008) find certain dialogue acts are important cues for detecting decisions in Multi-party dialogue. In Ranganath et al. (2009), dialogue acts are used as important features for flirt detection.

Automatic dialogue act recognition is still an active research topic. The conventional approach is to train one generic classifier using a large corpus of annotated utterances. One aspect that makes it so challenging is that people can express the same idea (or speech act) using a very different set of spoken words. Even more, people can mean different things with the exact same spoken words. These idiosyncratic differences in dialogue acts make the learning of generic classifiers extremely challenging. Luckily, in many applications such as face-to-face meetings or tele-immersion, we have access to archives of previous interactions with the same participants. From these archives, a small subset of spoken utterances can be efficiently annotated. As we will later show in our experiments, even a small number of annotated utterances can make a significant difference.

In this paper, we propose a new approach for dialogue act recognition based on reweighted domain adaptation which effectively balance the influence of speaker specific and other speakers' data. By treating each speaker as one domain, we point out the connection between training speaker specific dialogue act classifier and supervised domain adaptation problem. We analyze idiosyncracy in dialogue verbal acts by qualitatively studying the differences and conflicts among speakers and by quantitively comparing speaker-specific models. We present an extensive set of experiments studying the effect of speaker adaptation on dialogue act recogntion in multi-party meetings using the ICSI-MRDA dataset (Shriberg, 2004).

[1] This paper is an extended version of a poster presented at SemDial 2011, with new experiments and deeper analysis.

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 118–125,
Seoul, South Korea, 5-6 July 2012.

The following section presents related work on dialogue act recognition and domain adaptation. Section 3 describes the ICSI-MRDA (Shriberg, 2004) dataset which is used in all our experiments. Section 4 analyze idiosyncracy in dialogue acts, both qualitatively and quantitatively. Section 5 explains our reweighting-based speaker adaptation algorithm. Section 6 contains all experiments to prove the applicability of speaker adaptation to dialogue act recognition. Finally, inspired by the promising results, Section 8 describes some future directions.

2 Previous Work

Automatic dialogue act recognition has been an important problem in the past decades. Different dialogue act labeling standards and datasets have been provided, including Switchboard-DAMSL (Stolcke et al., 2000), ICSI-MRDA (Shriberg, 2004) and AMI (Carletta, 2007). Stolcke et al (2000) is one of the first work using machine learning technique (HMM) to automatically segment and recognize dialogue acts. Rangarajan et al. (2009) demonstrated well-designed prosodic n-gram features are very helpful for Dialogue Act recognition in Maximum Entropy model. And Ang et al (2005) explored joint segmentation and dialogue act classification for speech from ICSI.

Domain adaptation is a popular problem in natural language processing community due to the sparsity of labeled data. Jiang (Jiang, 2007) breaks the analysis of domain adaptation problem into distributional differences in instances and classification functions between source and target data. In Daume's work (2007) several domain adaptation algorithms are described. Our speaker adaptation algorithm is inspired by the reweighting-based adaptation algorithm introduced in this paper.

Recently, dialogue act adaptation has been getting a lot of attention. Tur et al. (2006) successfully use Switchboard-DAMSL to help dialogue act recognition in ICSI-MRDA. Promising results have been obtained by using a regression model to combine the model weights obtained by training on Switchboard-DAMSL and ICSI-MRDA respectively. Following the work by Tur et al. (2006), Guz et al. (2009) further studied the effectiveness of dialogue act domain adaptation in cascaded dialogue act segmentation and recognition system, their results prove adaptation in the intermediate step (segmentation) are also very helpful for the final output (recognition). Jeong et al (2009) use semi-supervised boosting algorithm to leverage labeled data from Switchboard-DAMSL and ICSI-MRDA to help dialogue act recognition in email and forums. Margolis et.al (2010) use a structural correspondence learning technique to adapt dialogue act recognition on automatic translated Spanish genre with the help of Switchboard-DAMSL and ICSI-MRDA. Kolar et al. (2007) explores the difference among speakers for dialogue act segmentation in ICSI-MRDA dataset. Similar to the approach taken in Tur et al. (2006), adaptation is performed through the combination of generic speaker independent Language Model and other speakers' Language Model. Significant improvements have been obtained for most of the selected speakers.

All these previous papers focused on adapting dialogue act models between domains and did not address the person-specific adaptation. The only exception was Kolar et al. (2007) who explored speaker-specific dialogue act segmentation. To our knowledge, this paper is the first work to analyze the effectiveness of speaker adaptation for dialogue act recognition.

3 ICSI-MRDA Corpus

Different Dialogue Act labeling standards and datasets have been provided in recent years, including Switchboard-DAMSL (Stolcke et al., 2000), ICSI-MRDA (Shriberg, 2004) and AMI (Carletta, 2007). ICSI-MRDA is the dataset for our experiments because many of its meetings contain the same speakers, thus making it more suitable for our speaker adaptation study. The tagset in ICSI-MRDA is adapted from DAMSL standard (damsl, 1997) by allowing multiple tags per dialogue act. Each dialogue act in ICSI-MRDA has one general tag and multiple specific tags.

ICSI-MRDA consists of 75 meetings, each roughly an hour long. There are five categories of meetings (three of which we are actively using in our experiments) : *Bed* is about the discussion of natural language processing and neural theories of language, *Bmr* is for the discussion on ICSI meeting corpus, *Bro* is on speech recognition topics and *Bns*

ID	Tag	Type	Nb. Meetings	Nb. DAs
1	mn015	Bed	15	6228
2	me010	Bed	11	5309
3	me013	Bmr	25	9753
4	mn017	Bmr	15	4059
5	fe016	Bmr	18	5500
6	me018	Bro	20	4263
7	me013	Bro	22	11928

Table 1: The 7 speakers from ICSI-MRDA dataset used in our experiments. The table lists: the Speaker ID, original speaker tag, the type of meeting selected for this speaker, the number of meetings this speaker participated and the total number of dialogue acts by this speaker.

is about network and architecture. The last category is *varies* which contains all other topics.

From these 75 meetings, there are 53 unique speakers in total, and an average of about 6 speakers per meeting. 7 speakers[2] having more than $4,000$ dialogue acts are selected for our adaptation experiments. Table 1 shows the details of our 7 selected speakers. From the word transcriptions, we created an extended list of linguistic features per utterance. From the 7 selected speakers, we computed 14653 unigram features, 158884 bigram features and 400025 trigram features.

Following the work of Shriberg et al. (2004), we use the 5 general tags in our experiments:

- *Disruption* indicates the current Dialogue Act is interrupted.
- *Back Channel* are utterances which are not made directly by a speaker as a response and do not function in a way that elicits a response either.
- *Floor Mechanism* are dialogue acts for grabbing or maintaining the floor.
- *Question* is for eliciting listener feed back.
- And finally, unless an utterance is completely indecipherable or else can be further described by a general tag, then its default status is *Statement*.

Our dataset consisted of 47040 dialogue acts. The distribution of Dialogue Act is shown in Table 2.

[2]speaker me013 is split into me013-Bmr and me013-Bro to avoid the difference introduced by meeting types.

Tag	proportion
Disruption	14.73%
Back Channel	10.20%
Floor Mechanism	12.40%
Question	7.20%
Statement	55.46%

Table 2: Distribution of dialogue acts in our dataset.

4 Idiosyncrasy in Dialogue Acts

Our goal is to create a dialogue act recognition algorithm that can adapt to specific speakers. Some important questions must be studied before creating such algorithm. The first obvious one is: do speakers really differ in their choice of words and associated dialogue acts? Do we really see a variability on how people express their dialogue intent? If the answers are yes, then we will expect that learning a dialogue act recognizer from speaker-specific utterances should always outperform a recognizer learned from someone else data. Section 4.1 presents a comparative experiment addressing these questions.

To better understand the results from this comparative experiment, we also performed a qualitative analysis presented in Section 4.2 where we look more closely at the differences between speakers. These two qualitative and quantitative analysis are building block for our adaptation algorithm presented in Section 5.

4.1 Speaker-Specific Recognizers

An important assumption when performing speaker adaptation (or more generally domain adaptation) is that data coming from the same speaker should be similar than data coming from another person. In other words, a recognizer trained on a speaker should perform better (when tested on the same person) than a recognizer trained on another speaker. We designed an experiment to test this hypothesis.

We learned 7 speaker-specific recognizers, one for each speaker (see Table 1). We then tested all these recognizers on new utterances from the same 7 speakers. We looked the recognition performance when (1) the recognizer was trained on the same person and (2) when the recognizer was trained on a different person. This experiments quantitatively

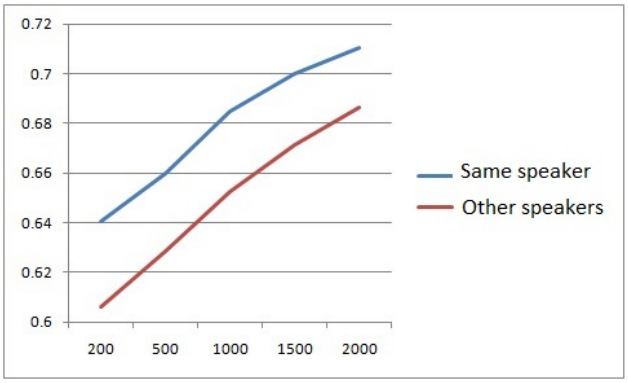

Figure 1: Effect of same-speaker data on dialogue act recognition. We compare two approaches: (1) when a recognizer is trained on the same person and tested on new utterances from the same person, and (2) when the recognizer was trained on another speaker (same test set). We vary the amount of training data to be 200, 500, 1000, 1500 and 2000 dialogue acts. In all cases, using speaker-specific recognizer outperforms recognizer from other speakers.

analyze the the difference among speakers. The experimental methodology used in this experiment is the same as the other experiments described in this paper (see Section 6). We use the Maximum Entropy model(MaxEnt) for all dialogue act recognizers (Ratnaparkhi, 1996). Please refer to Section 6.2 for more details about the experimental methodology.

Figure 1 compares the average performances when testing on the same speaker or on some other speaker. We vary the number of training data for each speaker to be 200, 500, 1000, 1500 and 2000 dialogue acts. For all five cases, the recognizers trained on the same speaker outperforms the average performance when using a recognizer from an other person. Thus speaker specific dialogue acts adaptation fits the assumption of domain adaptation problems.

4.2 Speakers Differences

To better understand the problem, we look more closely at the differences among speakers and their use of dialogue acts. We analyze the problem induced by speaker idiosyncrasy in dialogue acts. During our qualitative analysis of the ICSI-MRDA dataset, we identified three major differences explaining the performances observed in the previous sections: dialogue act conflicts, word distribution and dialogue act label distribution. We describe these three differences with some examples:

Conflicts: These differences happen when two speakers intended to express different meanings while speaking the exact same utterance. To examplify these conflicts, we computed mutual information between a specific utterance and all dialogue act labels. We find interesting examples where for exemple the word *right* is the most important cue for dialogue act *question* when spoken by me013-Bmr, while *right* is also an important cue for dialogue act *back-channel* for speaker me010-Bed. These examples suggest that conflicts exist among speakers and simply trying to learn one generic model may not be able to handle these conflicts. The generic model will learn what most people mean with this utterance, which may be the wrong prediction for our specific speaker.

Word distribution: People have their own vocabulary. Although many words are the same, how often one person use each word will vary. Although we may not have direct conflict here, the problem can also be serious. The learning algorithm may misleadingly focus on optimizing the weights for certain words which are not important(e.g., words that occur more often in other speakers' dialogue acts than his/her own) while under-estimating the important words for this speaker. This observation suggests that our adaptation should take into account word distribution.

Label Distribution: Another interesting observation is to look at the distribution of dialogue act labels for different speakers. Table 2 shows the average distribution over all 7 speakers. When looking more closely at each speaker, we find some interesting differences. For example, speaker 1 made statements 61% of the time while speaker 4 made 49% of the time. While this difference may not look significant, these changes can definitely affect the recognition performance. So the adaptation model should also take into account the dialogue act label distribution.

5 Reweighted Speaker Adaptation

Based on the observations described in the previous sections, we implement a simple reweighting-based

domain adaptation algorithm mentioned in (Daume, 2007) based on Maximum Entropy model (MaxEnt) (Ratnaparkhi, 1996). MaxEnt model is a popular and efficient discriminative model which can effectively accommodate large numbers of features. All the unigram, bigram and trigram features are used as input to the maxEnt model, the output is the dialogue act label. MaxEnt model maximizes the log conditional likelihood of all samples:

$$Loss = \sum_{1}^{N} log(p(y_n|x_n)) \quad (1)$$

where N is the number of samples for the training data. x_n represents the feature of the n_{th} sample and y_n is the label. The conditional likelihood is defined as

$$p(y|x) = exp(\sum_{i} \lambda_i f_i(x, y))/Z(x) \quad (2)$$

where $Z(x)$ is the normalization factor and $f_i(x, y)$ are the n-gram features described in Section 3.

When applied to our problem of speaker adaptation, the *reweighting adaptation model* can be formally defined as

$$Loss = w \sum_{n=1}^{S} log(p(y_n|x_n)) + \sum_{m=1}^{O} log(p(y_m|x_m)) \quad (3)$$

where S is the number of labeled speaker-specific dialogue acts, O is the number for other speakers' labeled dialogue acts. For each speaker, we train one speaker-specific classifier by varying the distribution of training data. We reweight the importance of speaker specific dialogue acts versus other speakers' labeled dialogue acts in the training data. The optimal weight parameter w is automatically estimated through validation.

It is worth mentioning a specific instance of the reweighting adaptation algorithm. When w is set to 1, the reweighting adaptation algorithm is equivalent to simply training a MaxEnt model by putting the speaker-specific and generic data samples together as training data. In our experiments, we will compare the reweighting adaptation approach with this simpler approach, referred as *constant adaptation*.

6 Experiments

Our goal is to get one model specifically adapted for each speaker. We first describes 4 different approaches to be compared in the experiments, and section 6.2 explains our experimental methodology.

6.1 4 Approaches

In these experiments, we compare our approach, called reweighted adaptation, with three more conventional approaches: speaker-specific only, Generic and Constant adaptation.

- ***Speaker Specific Only*** For this approach, we train the dialogue act recognizer using training sentences from the same speaker used during testing.
- ***Generic*** In this case, we train the dialogue act recognizer using utterances from all speakers other than the speaker used during testing.
- ***Constant Adaptation*** For this approach, we train the dialogue act recognizer using all speakers, including the speaker who will later be used for testing. All utterances have the same weight in this case.
- ***Reweighted Adaptation*** This is our proposed approach. As described in Section 5, we train our dialogue act recognizer using all speakers but reweight the utterances from the speaker who will later be used for testing.

6.2 Methodology

In all the following experiments we use MaxEnt models as defined in Section 5. $L2$ regularization is used for MaxEnt to avoid overfitting. The optimal regularization parameter was automatically selected during validation. The following regularization parameters were used: 0.01, 0.1, 1, 10 , 100, 1000 and 0 (no regularization). All the unigram, bigram and trigram features are used in the maxEnt model. The labels are the five dialogue act tags described in Section 3.

All experiments were performed using hold-out testing and hold-out validation. Both validation and test sets consisted of 1000 dialogue acts. The training sets contained only utterances from meetings that were not in the validation set of test set.

Train Data	200	500	1000	1500	2000
Speaker-specific Only	64.07	65.99	68.51	69.99	71.06
Constant adaptation model	76.81	76.96	77.00	77.23	77.53
Our reweighted adaptation model	78.17	78.29	78.67	78.74	78.47

Table 3: Average results among all 7 speakers when train with different combinations of speaker specific data and other speakers' data. The number of speaker specific data is varied from 200, 500, 1000, 1500 to 2000.

In many of our experiments, we analyzed the effect of training set size on the recognition performance. The speaker-specific data size varied from 200, 500, 1000, 1500 and 2000 dialogue acts respectively. When training our reweighting adaptation algorithm described in Section 5, we used the following weights: 10, 30, 50, 75, and 100. The optimal weight factor was selected automatically during validation.

7 Results

In this section we present our approaches to study the importance of speaker adaptation for dialogue act recognition. All following results are calculated based on the overall tag accuracies. We designed three series of experiments for this study:

- Generic Recognizer (Section 7.1)
- Sparsity in speaker-specific data (Section 7.2)
- Effectiveness of Constant Adaptation (Section 7.3)
- Performance of the reweighting algorithm (Section 7.4)

7.1 Generic Recognizer

The first result we get is on average, for each speaker when we use *all other speaker's* data for training, then test on speaker- specific test data. The performance of this generic recognizer is 76.76% is the baseline we try to improve when adding speaker-specific data into consideration. [3]

[3]The performance of our generic model is comparable to the results from Ang et al (2005) when you take into consideration that we used only 47,040 dialogue acts in our experiments (i.e., dialogue acts from our 7 speakers) which is a small fraction compared with Ang et al (2005) .

7.2 Sparsity of speaker-specific data

A second result is the performance when only using speaker-specific data. The row *Speaker Specific Only* in Table 3 shows the average results among all speakers when for each speaker, we train using only data from the same speaker. The number of speaker-specific training data we tried are 200, 500, 1000, 1500, and 2000 respectively. Even with 2000 speaker-specific dialogue acts for training, the best accuracy is 71.06% which is lower than 76.76% when using generic recognizer. Given the challenge in getting 2000 speaker-specific annotated dialogue acts, we are looking at a different approach where we need less speaker-specific data.

7.3 Results of Constant Adaptation

The most straightforward way to combine other speakers' data is to directly add them with speaker-specific data as train. We refer to this approach as constant adaptation. The row *Constant Adaptation* in Table 3 shows the average results among all speakers when for each speaker, we combine the speaker-specific data directly with the *all other speaker's* data. In our experiments, we varied the amount of speaker-specific data included to be 200, 500, 1000, 1500, and 2000 respectively. For all 7 speakers, the performance can always been improved by including speaker-specific data with all other speakers' data for training. Furthermore, the more speaker specific data added, the better performance we get.

7.4 Results of Reweighting Algorithm

Finally, in this section we describe the results for a simple adaptation algorithm based on reweighting, as described in Section 5. Following the same methodology as previous experiments, we vary the amount of speaker-specific data to be 200, 500, 1000, 1500 and 2000. The best reweighting factor is selected through validation on speaker-specific validation data described in section 6.2. The results of all 7 speakers from Reweighting algorithm when we vary the amount of speaker-specific data are shown in Figure 3.

We analyze the influence of the weighting factor on our speaker adaptation by plotting the recognition performance for different weights. Figure 4 il-

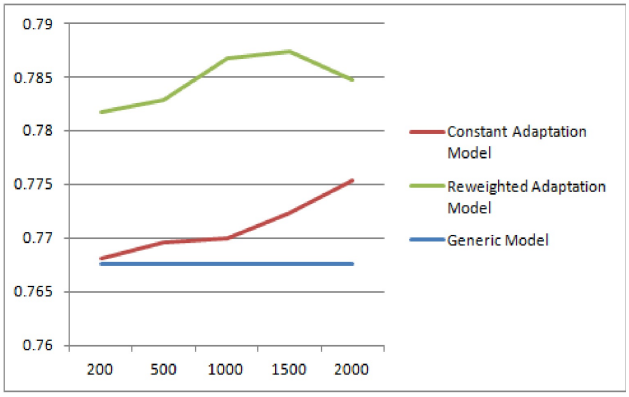

Figure 2: The average results among all 7 speakers when train with different combinations of speaker specific data and other speakers' data are displayed. In both Constant adaptation and Reweighted adaptation models the number of speaker specific data are varied from 200, 500, 1000, 1500 to 2000. In Generic model, only all other speakers' data are used for training data.

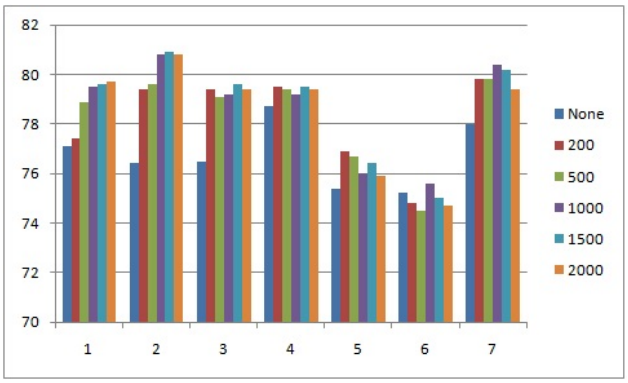

Figure 3: Reweighting algorithm for all 7 Individual Speakers when varying the amount of training data to be 0, 200, 500, 1000, 1500 and 2000.

lustrates the influence of the weight factor on three speaker adaptation cases: None, 500 and 2000. In this case, None represent the Constant Adaptation. We observe the following trend: with more speaker-specific data, the optimal reweighting factor is also lower. This confirms that our reweighting algorithm finds the right balance between speaker-specific data and generic data.

Figure 2 and the row *Reweighted Adaptation* from Table 3 shows the effectiveness of reweighting algorithm. Results shows that even this simple algorithm can efficiently balance the influence of speaker specific data and other speakers' data and

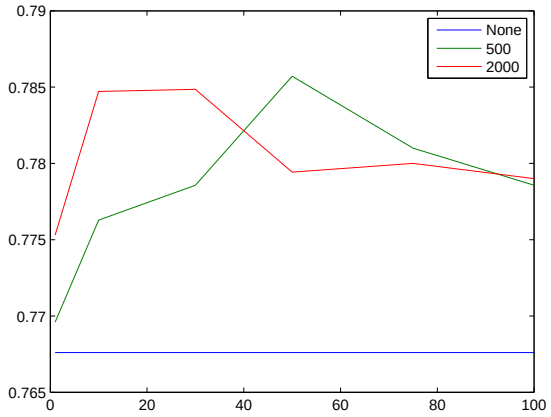

Figure 4: Average results of Reweighting among all 7 speakers when the amount of speaker specific data is 0, 500, 2000

give significantly improved results. And most surprisingly, even with only 200 speaker specific data the reweighting algorithm can give very promising results.

8 Conclusion

In this work we analyze the effectiveness of speaker adaptation for dialogue act recognition. A simple reweighting algorithm is shown to give promising improvement on several baseline algorithms even with only 200 speaker-specific dialogue acts. This paper is a first step toward automatic adaptation for dialogue act recognition. Inspired by the promising results from the simple reweighting algorithm, we plan to evaluate other domain adaptation techniques such as Daume's feature-based approach (2007). It will also be interesting to consider the unlabeled data from each speaker when performing dialogue act recognition.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1118018 and the U.S. Army Research, Development, and Engineering Command (RDECOM). The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Jeremy Ang, Yang Liu, Elizabeth Shriberg 2005. Automatic Dialog Act Segmentation and Classification in Multiparty Meetings. *ICASSP*.

Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41(2): 181-190

Mark Core and James Allen. 1997. Working Notes: AAAI Fall Symposium. *HLT-NAACL SIGDIAL Workshop*.

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

Raquel Fernandez, Matthew Frampton, Patrick Ehlen, Matthew Purver and Stanley Peters. 2008. Modelling and Detecting Decisions in Multi-Party Dialogue. *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*.

Umit Guz, Gokhan Tur, Dilek Hakkani-Tur, and Sebastien Cuendet. 2009. Cascaded model adaptation for dialog act segmentation and tagging. *Computer Speech & Language*, 24(2):289–306.

Minwoo Jeong, Chin-Yew Lin and Gary Lee. 2009. Semi-supervised speech act recognition in emails and forums. *The 2009 Conference on Empirical Methods on Natural Language Processing*.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

Jachym Kolar, Yang Liu, and Elizabeth Shriberg. 2007. Speaker Adaptation of Language Models for Automatic Dialog Act Segmentation of Meetings. *Interspeech*, 339–373.

Anna Margolis, Karen Livescu, Mari Ostendorf. 2010. Semi-supervised domain adaptation for automatic dialog act tagging. *ACL 2010 Workshop on Domain Adaptation for Natural Language Processing*.

Rajesh Ranganath, Dan Jurafsky, and Dan McFarland. 2009. It's Not You, it's Me: Detecting Flirting and its Misperception in Speed-Dates. *The 2009 Conference on Empirical Methods on Natural Language Processing*.

Vivek Rangarajan, Srinivas Bangaloreb and Shrikanth Narayanana. 2009. Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. *Computer Speech and Language*, 23(4): 407-422

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang and Hannah Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. *HLT-NAACL SIGDIAL Workshop*.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol V. Ess-dykema and Marie Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26:339-373.

Gokhan Tur, Umit Guz and Dilek Hakkani-Tur. 2006. Model Adaptation For Dialogue Act Tagging. *Spoken Language Technology Workshop.*