SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, pages 188–189.
Los Angeles, California, 21–23 September 2011.

Towards Speaker Adaptation for Dialogue Act Recognition

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

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

1 Introduction

Dialogue act labels are being used to represent a higher level intention of utterances during human conversation (Stolcke et al., 2000). 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 (Stolcke et al., 2000). One aspect that makes it so challenging is that people can express the same intentions using a very different set of spoken words. Imagine how different the vocabulary used by a native English speaker or a foreigner can be. Even more, people can have different intentions when using 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 differences in the dialogue act recognition performance.

In this paper, we propose a new approach for dialogue act recognition based on reweighted domain adaptation inspired by Daume's work (2007) which effectively balance the influence of speaker specific and other speakers' data. We present a preliminary set of experiments studying the effect of speaker adaptation on dialogue act recognition in multi-party meetings using the ICSI-MRDA dataset (Shriberg, 2004). To our knowledge, this paper is the first work to analyze the effectiveness of speaker adaptation for automatic dialogue act recognition.

2 Balanced Adaptation

Different people may have different patterns during conversation, thus learning a single generic model for all people is usually not optimal in dialogue act recognition task. In this work, for each speaker, we construct a balanced speaker adapted classifier based on a simple reweighting-based domain adaptation algorithm from Daume (2007).

Model parameters are learned through the minimization of the loss function defined as the sum of log likelihood on speaker specific data and other speakers' data

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

S is a set containing all labeled speaker-specific dialogue acts, O is a set containing all other speakers' labeled dialogue acts. w is for balancing the importance of speaker specific data versus other speaker's data. x_n and x_m are the utterances features, y_n and y_m are the dialogue act labels, $p(y_n|x_n)$ and $p(y_m|x_m)$ are defined as

$$p(y|x) = exp(\sum_i \lambda_i f_i(x, y))/Z(x). \quad (2)$$

3 Experiments

In this paper, we selected the ICSI-MRDA dataset (Shriberg, 2004) for our experiments because many of its meetings contain the same speakers, thus making it better suited for our speaker adaptation study. ICSI-MRDA consists of

Models	200	500	1000	1500	2000
Generic	76.76%				
Speaker only	64.07%	65.99%	68.51%	69.99%	71.06%
Simple speaker adaptation	76.81%	76.96%	77.00%	77.23%	77.53%
balanced speaker adaptation	**78.17%**	**78.29%**	**78.67%**	**78.74%**	**78.47%**

Table 1: Average results among all 7 speakers when train with different combinations of speaker specific data and other speakers' data and vary the amount of training data to be 200, 500, 1000, 1500 and 2000.

75 meetings, each roughly an hour long. From these 75 meetings, we selected for our experiments 7 speakers who participated in at least 10 meetings and spoke more than $4,000$ dialogue acts. From the utterance transcriptions, we computed $14,653$ unigram features, $158,884$ bigram features and $400,025$ trigram features. Following the work of Shriberg et al. (2004), we used the 5 general tags *Disruption*(14.7%), *Back Channel(*10.20%*)*, *Floor Mechanism*(12.40%), *Question*(7.20%) and *Statement*(55.46%) as labels. The total number of dialogue acts for all 7 speakers was $47,040$.

All experiments were performed using hold-out testing and hold-out validation. Both validation and testing sets consisted of 1000 dialogue acts from meetings not in the training set. In 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 balanced adaptation algorithm described in Section 2, we validated the balance factor w using the following values: 10, 30, 50, 75 and 100. The optimal balance factor w was selected automatically during validation. The following four experiments are intended to prove the effectiveness of speaker balanced adaptation. Their respective results are listed in Table 1.

1. **Generic** represents the conventional method where a large corpus is used to train the recognizer and then tested on a new person who is not part of the training. The average accuracy over the 7 participants is 76.7%.

2. **Speaker Only** represents the approach where we train a recognizer using only one person data and test on spoken utterances from the same person. We show in Table 1 the average accuracy over our 7 participants for different size of training sets. Even with 2000 speaker-specific dialogue acts for training, the best accuracy is 71.06% which is much lower than 76.76% from the generic recognizer. Given the challenge in labeling 2000 speaker-specific annotated dialogue acts, we are looking at a different approach where we need less speaker-specific data.

3. **Simple speaker adaptation** represents the approach where the training set consists of all the generic utterances(from other participants) and a few utterances from the speaker of interest(same speaker used during testing). This approach is equivalent to keeping a balance factor w of 1 in equation (1). Results showing that for all 7 speakers, the accuracy always improve when including speaker-specific data with all other speakers' data for training.

4. **Balanced speaker adaptation** shows the results for balanced adaptation algorithm described in section 2. This algorithm shows significant improvement over all the other approaches in Table 1 even with only 200 speaker-specific dialogue acts. These results show that with even a simple adaptation algorithm we can improve the automatic dialogue act recognition.

References

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.

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

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