CO-ADAPTATION: ADAPTIVE CO-TRAINING FOR SEMI-SUPERVISED LEARNING

Gokhan Tur

SRI International,
Speech Technology and Research Lab
Menlo Park, CA, 94025
gokhan@speech.sri.com

ABSTRACT

Inspired by popular co-training and domain adaptation methods, we propose a co-adaptation algorithm. The goal is improving the performance of a dialog act segmentation model by exploiting the vast amount of unlabeled data. This task provides a nice framework for multiview learning, as it has been shown that lexical and prosodic features provide complementary information. Instead of simply adding machine-labeled data to the set of manually labeled data, co-adaptation technique adapts the existing models. While both co-training and domain adaptation techniques have been employed for dialog act segmentation, our experiments show that the proposed co-adaptation algorithm results in significantly better performance.

Index Terms— co-adaptation, co-training, semi-supervised learning, domain adaptation

1. INTRODUCTION

Recent advances in data-driven speech and language processing techniques combined with discriminative machine learning algorithms, such as support vector machines or boosting, enable us to build high-performance, robust, portable, statistical models. However, data-driven classifiers are trained using large amounts of in-domain task data that is usually transcribed and then labeled by humans, an expensive and laborious process. To this end, in the literature, semi-supervised learning techniques such as self-training or co-training [1], and supervised or unsupervised domain adaptation techniques [2] have been proposed.

The aim of semi-supervised learning is to exploit the smaller amount of existing labeled data and larger amount of unlabeled data from a given domain. On the other hand, domain adaptation tries to build a new adapted model by using the existing out-of-domain model together with the small amount of labeled in-domain data. Since the in-domain data and the out-of-domain data do not usually share the same distribution, the out-of-domain classification models are typically adapted before they are employed.

The most common method of semi-supervised learning is self-training in which the given model estimates the classes for the unlabeled portion of the data. Then the examples that are classified automatically are added to the training set, the model is retrained, and the whole process is iterated [3]. Self-training has been applied to many speech and language processing tasks. In many studies, it has been shown that co-training outperforms self-training. First suggested by [1], co-training requires multiple views of the given task. Then, models trained from each of these views can provide machine-annotated data for the other views.

In the liteature, surprisingly, there are not many studies that naturally combine semi-supervised learning and domain adaptation. In other words, the machine-annotated data is treated in the same way as human annotated data. However, these typically have much different distributions in terms of classes or features. In our previous work, for call-type classification, we performed model adaptation while adding automatically labeled data for semi-supervised learning using self-training [4].

In this paper, we propose to extend the co-training algorithm with model adaptation techniques. Instead of simply adding machine-labeled data to the set of manually labeled data, we adapt the existing model using the machine-labeled data. The resulting algorithm, which we call *co-adaptation*, is then tested on a well-defined speech processing task of dialog act segmentation [5].

In the next section we describe the dialog act segmentation task and our approach. Sections 3 and 4 provide the co-training and adaptation algorithms established in the literature. Section 5 presents the proposed co-adaptation approach. In Section 6, we present the experiments and results using the ICSI MRDA Multiparty Meeting Corpus.

2. DIALOG ACT SEGMENTATION

Dialog acts (DAs) are basic building blocks 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 questions or backchannels [6]. Dialog acts are designed to be task independent, their main goal being to provide a basis for further discourse analysis and understanding. There are a number of predefined dialog act sets in the literature, such as Dialog Act Markup in Several Layers (DAMSL) [7] and MRDA [8].

Typically, dialog act tagging is performed on the automatic speech recognition (ASR) output. Since most ASR outputs lack typographic cues such as sentence and paragraph boundaries, an intermediate segmentation step is necessary. This task, known as *sentence unit segmentation* or *dialog act segmentation*, aims at deciding whether a particular word boundary marks the end of a dialog act unit.

Dialog act segmentation is generally framed as a word boundary classification problem. For DA segmentation [9] used a method that combines hidden Markov models (HMMs) with N-gram language models containing words and dialog act boundaries associated with them. [10] provides an overview of different classification algorithms (boosting, hidden-event language model, maximum entropy and decision trees) applied to the dialog act segmentation for multilingual broadcast news. Besides the type of classifier, the features have widely been studied; [9, 11] showed how prosodic features can benefit the dialog act segmentation task. Investigations on prosodic and lexical features in the context of phone conversation and broadcast news speech were presented in [11]. More recently, [12] studied

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

```
• Given a set $L$ of labeled and a set $U$ of
  unlabeled examples
• Loop for $k$ iterations
    – Train a classifier $H_1$ on view 1 of $L$
    – Train a classifier $H_2$ on view 2 of $L$
    – Allow $H_1$ to label $U$, add $n$ most
      confidently labeled samples to $L$
    – Allow $H_2$ to label $U$, add $n$ most
      confidently labeled samples to $L$
```

Fig. 1. The original co-training algorithm.

syntactic features for this task.

In this work, for dialog act segmentation, a sample (that is, a word boundary b_i, between words w_i and w_{i+1}) is represented by features containing lexical information (word n-grams) and prosodic information (the pause duration between the two words at the word boundary of interest, and various measures of the pitch and the energy of the voice of the speaker). With these features, we use the AdaBoost.MH algorithm, a member of the boosting family of classifiers, which has been shown to be among the best classifiers for the sentence segmentation task [10]. Boosting is an iterative procedure that builds a new weak learner h_t at each iteration. Every example of the training data set is assigned a weight. These weights are initialized uniformly and updated on each iteration so that the algorithm focuses on the examples that were wrongly classified on the previous iteration. At the end of the learning process, the weak learners used on each iteration t are linearly combined to form the classification function:

$$f(x,l) = \sum_{t=1}^{T} \alpha_t h_t(x,l)$$

with α_t the weight of the weak learner h_t and T the number of iterations of the algorithm. This algorithm can be seen as a procedure for finding a linear combination of base classifiers that attempts to minimize a loss function, such as the logistic loss:

$$\sum_i \sum_l \ln(1 + e^{-Y_i[l] f(x_i,l)}).$$

More details on boosting can be found in [13].

3. CO-TRAINING ALGORITHM

The co-training algorithm was first proposed by Blum and Mitchell [1]. For using co-training, the features in the problem domain should naturally divide into two sets. Then the examples that are classified with high confidence scores with one view can be used as the training data of other views. For example, for web page classification, one view can be the text in them and another view can be the text in the hyperlinks pointing to those web pages. Similarly for dialog act segmentation, lexical and prosodic features provide two naturally divided views.

While the original co-training algorithm (provided in Figure 1 for two views) is widely popular and effective in most tasks, researchers have extended it in a number of ways. [14] combined co-training and the expectation maximization (EM) algorithms, coming up with the Co-EM algorithm that can be seen as the probabilistic version of co-training. In this version, instead of choosing n samples, all data is added to L with classification confidence. A naive Bayes classifier that can easily accommodate this extension is employed.

```
• Given a set $L$ of labeled and a set $U$ of
  unlabeled examples
• Loop for $k$ iterations
    – Train a classifier $H_1$ on view 1 of $L$
    – Train a classifier $H_2$ on view 2 of $L$
    – Allow $H_1$ and $H_2$ to label $U$
    – Add $n$ samples to $L$ which are
      confidently labeled by $H_1$ and
      unconfidently labeled by $H_2$
    – Add $n$ samples to $L$ which are
      confidently labeled by $H_2$ and
      unconfidently labeled by $H_1$
```

Fig. 2. The disagreement-based co-training algorithm.

Later, [15] and [16] almost simultaneously suggested considering classifiers trained from other views while choosing n samples to label automatically. Two methods called *agreement* (or max-t-max-s) and *disagreement* (or max-t-min-s) happen to outperform the original co-training method, and the latter one resulted in better performance. The agreement-based approach chooses samples that are labeled by all the classifiers trained with different views similarly with high confidence. The disagreement method chooses the samples labeled by the classifiers trained from other views with high confidence, but not the one in question. For example, for dialog act segmentation, if the prosodic model classifies an example with high confidence and the lexical model is undecided, this sample is added to the training set of the lexical model, and vice versa. The disagreement-based co-training algorithm with two views is more formally shown in Figure 2. One thing to note is that, in these extended methods, the training sets of multiple views are kept separately, as turned out to be superior in performance.

In our previous work, these co-training methods were analyzed and compared for dialog act segmentation using lexical and prosodic models [15]. Using the ICSI Meeting Corpus we have reported significant improvements by performing co-training.

4. MODEL ADAPTATION

Model adaptation has been extensively studied in speech processing, notably for acoustic and language modeling. Two very popular adaptation approaches are maximum likelihood linear regression (MLLR) [17] and maximum *a posteriori* (MAP) adaptation [18]. For language processing tasks, the most common approach is model interpolation (e.g., [19]). In model interpolation, an out-of-domain model θ_{OOD} is interpolated with an in-domain model θ_{ID} to form an adapted model $\hat{\theta}$:

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

For boosting, [20] explored model adaptation via changing the loss function during training:

$$\sum_j \left[\ln\left(1 + e^{-y_j f(x_j)} + \eta KL\left(\pi_+(x_j) \| \sigma(f(x_j))\right)\right)\right]$$

where KL is the binary relative entropy (or Kullback-Leibler divergence), $1 + e^{-y_j f(x_j)}$ and $\pi_+(x_j)$ the logistic loss function and

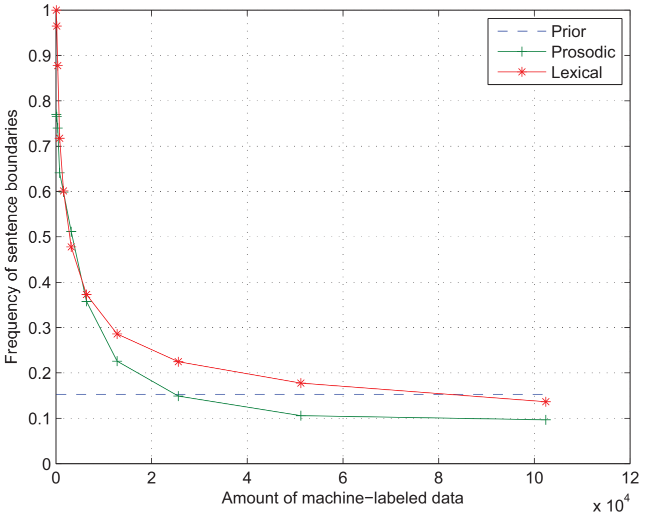

Fig. 3. Frequency of the sentence boundaries after the first iteration of co-training for prosodic and lexical models.

```
• Given a set $L$ of labeled and a set $U$ of
  unlabeled examples
• Train a classifier $H_1$ on view 1 of $L$
• Train a classifier $H_2$ on view 2 of $L$
• Loop for $k$ iterations

    – Allow $H_1$ and $H_2$ to label $U$
    – Adapt $H_1$ using $n$ samples which
      are confidently labeled by $H_2$ and
      unconfidently labeled by $H_1$
    – Adapt $H_2$ using $n$ samples which
      are confidently labeled by $H_1$ and
      unconfidently labeled by $H_2$
```

Fig. 4. The co-adaptation algorithm.

	Train	Dev	Test
Number of Sentences	82,435	16,720	16,397
Number of Words	538,956	110,851	101,510
Average Sentence Length	6.53	6.62	6.19

Table 1. Data characteristics used in the experiments.

prior distribution of the out-of-domain model, and $\sigma(f(x_j))$ the distribution of the in-domain model. This is the same as minimizing a weighted sum of the logistic loss function and the binary relative entropy of the prior probabilities of both models. The weight η is optimized using a held-out set.

In our previous work, these adaptation methods were analyzed and compared for dialog act segmentation [21]. The task was adapting the models trained from human/human conversations (e.g., Switchboard corpus) to multiparty meetings (e.g., ICSI Meeting Corpus). We have reported significant improvements by performing domain adaptation using almost all the alternative techniques except simple data concatenation, which is equivalent to uniformly weighted model interpolation.

5. CO-ADAPTATION ALGORITHM

The main idea of co-adaptation is similar to co-training; however, instead of simply adding machine-labeled data to the set of manually labeled data, we adapt the existing model using the machine-labeled data with some weight tuned using a held-out set. This is inspired by the fact that while doing model adaptation, unweighted adaptation or data concatenation is the worst possible approach [21].

Our co-training experiments show an interesting phenomenon about the examples selected at each iteration of co-training: their distribution is significantly different from the labeled set. This is actually intuitive since these are the hard examples that need special attention. Check Figure 3 that shows the frequency of the sentence boundaries after the first iteration of disagreement-based co-training. Since there are many hard to classify examples by one view which are easy for the other to detect the sentence boundary class, the majority of the selected examples are the sentence-final words. While one may enforce a selection mechanism enforcing a given distribution, this is still not a solution since the features are completely different as well. For example, the data added to the lexical model has an average pause duration of 15 frames, while the prior is 1.5 frames since most words do not have a pause duration following them. The feature and class distributions of the selected samples are very different, as if they are from a different domain, hence the use of domain adaptation methods.

The algorithm is shown more formally in Figure 4. This is based on the disagreement-based co-training algorithm, although it is applicable to any specific implementation or extension of the generic co-training algorithm.

Note that one can use any model adaptation technique, depending on the task and classification method used. In our experiments, since we built the dialog act segmentation models using boosting, the natural choice was using boosting adaptation as described above.

Note that this is different from the co-adaptation algorithm proposed by Christoudias *et al.* [22]. Their algorithm aims at exploiting out-of-domain data hence extending domain adaptation with co-training. More specifically, using the labeled out-of-domain data, they automatically label a small amount of in-domain data, which is then used to build the seed model for co-training. In the co-adaptation algorithm proposed in this paper, all data is in-domain, and a small amount is assumed to be manually labeled.

6. EXPERIMENTS AND RESULTS

We performed controlled experiments for analyzing the effectiveness of the co-adaptation method using the ICSI Meeting Corpus [8]. We drew learning curves by changing the size of the manually labeled data and using the rest of the data to evaluate the co-adaptation approach. We used 51 meetings for training, 11 meetings for tuning, and 11 meetings for testing as in [23]. All the experiments were done using manually trancribed data in order not to deal with automatic speech recognition noise. We performed our tests using the Boostexter tool [24]. The data characteristics are shown in Table 1. We repeated all the experiments three times, by shuffling the training data and averaged the performance figures. In these experiments we tuned the system parameters such as the adaptation weight using the held-out set.

As multiple views of the data, as mentioned above, we used

	1000 sentences		*2500 sentences*	
	Lexical	Prosodic	Lexical	Prosodic
Baseline	65.47	62.23	62.85	60.93
Co-Training	62.78	59.81	60.41	58.30
Co-Adaptation	**58.57**	**58.77**	**57.64**	**57.94**

Table 2. NIST error rates comparing co-training and co-adaptation for the lexical and prosodic models.

prosodic and lexical features extracted from the utterances. As the first set of experiments, we took only 1,000 and 2,500 sentences from the training set and employed co-training and co-adaptation techniques. Since in our previous work [15] we showed that disagreement-based co-training significantly outperforms self-training and conventional agreement-based co-training, we only compared the co-adaptation performance with the disagreement-based co-training method. Note that this is a very powerful method as proven for this and other tasks.

Table 2 presents our results in terms of NIST error rates. The NIST error rate is the ratio of the number of insertion and deletion errors for sentence boundaries made by the classifier to the number of reference sentence boundary classes. Therefore, if no boundaries are marked by sentence segmentation, it is 100%, but it can exceed 100%; the maximum error rate is the ratio of number of words to number of correct boundaries.

We provide results using a manually labeled seed set of 1,000 and 2,500 dialog act units. The table presents the performance of the lexical and prosodic models when co-training and co-adaptation methods are applied. Throughout the table, the co-adaptation method outperforms the co-training method, which is already significantly better than the baseline, reducing the NIST error rate up to 10.5% relative. After 5,000 sentences the effect of the semi-supervised learning starts to disappear consistent with our earlier work [15].

7. CONCLUSIONS

We have presented a method extending the well-known co-training algorithm using model adaptation techniques. The resulting algorithm, called *co-adaptation*, has been evaluated for the task of dialog act segmentation. Our results indicate that using co-adaptation resulted in as much as a five-fold saving for the amount of labeled data.

We believe that it is simply impossible to manually label all the available data. It is extremely important to work on lightly-supervised or unsupervised learning techniques for speech and language processing. Note that co-adaptation is a generic algorithm just as co-training and may be applied to other tasks in which co-training has resulted in significant improvements such as web page classification [1].

Our future work consists of experimenting with other established model adaptation methods, such as MAP, and employing co-adaptation when no in-domain data is available using a labeled out-of-domain data set.

Acknowledgments: This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) CALO program under Contract No. FA8750-07-D-0185. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA or the Department of Interior-National Business Center (DOI-NBC).

8. REFERENCES

[1] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” in *Proceedings of the COLT*, Madison, WI, July 1998.

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

[3] Kamal Nigam and Rayid Ghani, “Understanding the behaviour of co-training,” in *Proceedings of the Workshop on Text Mining at the Sixth ACM SIGKDD at the KDD*, 2000.

[4] G. Tur and D. Hakkani-Tür, “Exploiting unlabeled utterances for spoken language understanding,” in *Proceedings of the Eurospeech*, Geneva, Switzerland, September 2003.

[5] A. Stolcke and E. Shriberg, “Automatic linguistic segmentation of conversational speech,” in *Proceedings of ICSLP*, Philadelphia, PA, 1996.

[6] 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.

[7] 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.

[8] 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.

[9] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tur, “Prosody-based automatic segmentation of speech into sentences and topics,” *Speech Communication*, vol. 32, no. 1-2, pp. 127–154, 2000.

[10] M. Zimmerman, D. Hakkani-Tür, J. Fung, N. Mirghafori, L. Gottlieb, E. Shriberg, and Y. Liu, “The ICSI+ multilingual sentence segmentation system,” in *Proceedings of the ICSLP*, Pittsburg, PA, 2006.

[11] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, D. Hillard, M. Ostendorf, M. Tomalin, P. Woodland, and M. Harper, “Structural metadata research in the EARS program,” in *Proceedings of the ICASSP*, 2005.

[12] B. Roark, Y. Liu, M. Harper, R. Stewart, M. Lease, M. Snover, I. Shafran, B. Dorr, J. Hale, A. Krasnyanskaya, and L. Yung, “Reranking for sentence boundary detection in conversational speech,” in *Proceedings of the ICASSP*, Toulouse, France, 2006.

[13] R. E. Schapire, “The boosting approach to machine learning: An overview,” in *Proceedings of the MSRI Workshop on Nonlinear Estimation and Classification*, Berkeley, CA, March 2001.

[14] K. Nigam and R. Ghani, “Analyzing the effectiveness and applicability of co-training,” in *Proceedings of the CIKM*, McLean, VA, 2000.

[15] U. Guz, D. Hakkani-Tür, S. Cuendet, and G. Tur, “Co-training using prosodic and lexical information for sentence segmentation,” in *Proceedings of the INTERSPEECH*, Antwerp, Belgium, August 2007.

[16] W. Wang, Z. Huang, and M. Harper, “Semi-supervised learning for part-of-speech tagging of mandarin transcribed speech,” in *Proceedings of the ICASSP*, Honolulu, HI, 2007.

[17] M. J. F. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[18] J.-L. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[19] G. Tur, U. Guz, and D. Hakkani-Tür, “Model adaptation for dialog act tagging,” in *Proceedings of the IEEE/ACL SLT Workshop*, 2006.

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

[21] S. Cuendet, D. Hakkani-Tür, and G. Tur, “Model adaptation for sentence segmentation from speech,” in *Proceedings of the IEEE/ACL SLT Workshop*, Aruba, 2006.

[22] C. M. Christoudias, K. Saenko, L.-P. Morency, and T. Darrell, “Co-adaptation of audio-visual speech and gesture classifiers,” in *Proceedings of the ACM (ICMI)*, Banff, Alberta, Canada, 2006.

[23] 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.

[24] 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.