Domain Adaptation with Unlabeled Data for Dialog Act Tagging

Anna Margolis[1,2] **Karen Livescu**[2] **Mari Ostendorf**[1]
[1]Department of Electrical Engineering, University of Washington, Seattle, WA, USA.
[2]TTI-Chicago, Chicago, IL, USA.
amargoli@ee.washington.edu, klivescu@ttic.edu, mo@ee.washington.edu

Abstract

We investigate the classification of utterances into high-level dialog act categories using word-based features, under conditions where the train and test data differ by genre and/or language. We handle the cross-language cases with machine translation of the test utterances. We analyze and compare two feature-based approaches to using unlabeled data in adaptation: restriction to a shared feature set, and an implementation of Blitzer et al.'s Structural Correspondence Learning. Both methods lead to increased detection of backchannels in the cross-language cases by utilizing correlations between backchannel words and utterance length.

1 Introduction

Dialog act (or speech act) tagging aims to label abstract functions of utterances in conversations, such as Request, Floorgrab, or Statement; potential applications include automatic conversation analysis, punctuation transcription, and human-computer dialog systems. Although some applications require domain-specific tag sets, it is often useful to label utterances based on generic tags, and several tag sets have been developed for this purpose, e.g. DAMSL (Core and Allen, 1997). Many approaches to automatic dialog act (DA) tagging assume hand-labeled training data. However, when building a new system it may be difficult to find a labeled corpus that matches the target domain, or even the language. Even within the same language, speech from different domains can differ linguistically, and the same DA categories might be characterized by different cues. The domain characteristics (face-to-face vs. telephone, two-party vs. multi-party, informal vs. agenda-driven, familiar vs. stranger) can influence both the distribution of tags and word choice.

This work attempts to use unlabeled target domain data in order to improve cross-domain training performance, an approach referred to as both unsupervised and semi-supervised domain adaptation in the literature. We refer to the labeled training domain as the source domain. We compare two adaptation approaches: a simple one based on forcing the classifier to learn only on "shared" features that appear in both domains, and a more complex one based on Structural Correspondence Learning (SCL) from Blitzer et al. (2007). The shared feature approach has been investigated for adaptation in other tasks, e.g. Aue and Gamon (2005) for sentiment classification and Dredze et al. (2007) for parsing. SCL has been used successfully for sentiment classification and part-of-speech tagging (Blitzer et al., 2006); here we investigate its applicability to the DA classification task, using a multi-view learning implementation as suggested by Blitzer et al. (2009). In addition to analyzing these two methods on a novel task, we show an interesting comparison between them: in this setting, both methods turn out to have a similar effect caused by correlating cues for a particular DA class (Backchannel) with length.

We classify pre-segmented utterances based on their transcripts, and we consider only four high-level classes: Statement, Question, Backchannel, and Incomplete. Experiments are performed using all train/test pairs among three conversational speech corpora : the Meeting Recorder Dialog Act corpus (MRDA) (Shriberg et al., 2004), Switchboard DAMSL (Swbd) (Jurafsky et al., 1997), and the Spanish Callhome dialog act corpus (SpCH) (Levin et al., 1998). The first is multi-party, face-to-face meeting speech; the second is topic-prompted telephone speech between strangers; and the third is informal telephone speech between friends and family members. The first two are in English, while the third is in Spanish. When the source and target domains differ in language, we

Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing, ACL 2010, pages 45–52,
Uppsala, Sweden, 15 July 2010. ©2010 Association for Computational Linguistics

apply machine translation to the target domain to convert it to the language of the source domain.

2 Related Work

Automatic DA tagging across domain has been investigated by a handful of researchers. Webb and Liu (2008) investigated cross-corpus training between Swbd and another corpus consisting of task-oriented calls, although no adaptation was attempted. Similarly, Rosset et al. (2008) reported on recognition of task-oriented DA tags across domain and language (French to English) by using utterances that had been pre-processed to extract entities. Tur (2005) applied supervised model adaptation to intent classification across customer dialog systems, and Guz et al. (2010) applied supervised model adaptation methods for DA segmentation and classification on MRDA using labeled data from both MRDA and Swbd. Most similar to our work is that of Jeong et al. (2009), who compared two methods for semi-supervised adaptation, using Swbd/MRDA as the source training set and email or forums corpora as the target domains. Both methods were based on incorporating unlabeled target domain examples into training. Success has also been reported for self-training approaches on same-domain semi-supervised learning (Venkataraman et al., 2003; Tur et al., 2005). We are not aware of prior work on cross-lingual DA tagging via machine translation, although a translation approach has been employed for cross-lingual text classification and information retrieval, e.g. Bel et al. (2003).

In recent years there has been increasing interest in domain adaptation methods based on unlabeled target domain data. Several kinds of approaches have been proposed, including self-training (Roark and Bacchiani, 2003), instance weighting (Huang et al., 2007), change of feature representation (Pan et al., 2008), and clustering methods (Xing et al., 2007). SCL (Blitzer et al., 2006) is one feature representation approach that has been effective on certain high-dimensional NLP problems, including part-of-speech tagging and sentiment classification. SCL uses unlabeled data to learn feature projections that tie together source and target features via their correlations with features shared between domains. It first selects "pivot features" that are common in both domains; next, linear predictors for those features are learned on all the other features. Finally, singular value decomposition (SVD) is performed on the collection of learned linear predictors corresponding to different pivot features. Features that tend to get similar weights in predicting pivot features will be tied together in the SVD. By learning on the SVD dimensions, the source-trained classifier can put weight on target-only features.

3 Methods

Our four-class DA problem is similar to problems studied in other work, such as Tur et al. (2007) who used five classes (ours plus Floorgrab/hold). When defining a mapping from each corpus' tag set to the four high-level classes, our goal was to try to make the classes similarly defined across corpora. Note that the Incomplete category is defined in Swbd-DAMSL to include only utterances too short to determine their DA label (e.g., just a filler word). Thus, for our work the MRDA Incomplete category excludes utterances also tagged as Statement or Question; it includes those consisting of just a floor-grab, hold or filler word.

For classification we used an SVM with linear kernel, with L2 regularization and L1 loss, as implemented in the Liblinear package (Fan et al., 2008) which uses the one-vs.-rest configuration for multiclass classification. SVMs have been successful for supervised learning of DAs based on words and other features (Surendran and Levow, 2006; Liu, 2006). Features are derived from the hand transcripts, which are hand-segmented into DA units. Punctuation and capitalization are removed so that our setting corresponds to classification based on (perfect) speech recognition output. The features are counts of unigrams, bigrams, and trigrams that occur at least twice in the train set, including beginning/end-of-utterance tags (⟨s⟩, ⟨/s⟩), and a length feature (total number of words, z-normalized across the training set). Note that some previous work on DA tagging has used contextual features from surrounding utterances, or Markov models for the DA sequence. In addition, some work has used prosodic or other acoustic features. The work of Stolcke et al. (2000) found benefits to using Markov sequence models and prosodic features in addition to word features, but those benefits were relatively small, so for simplicity our experiments here use only word features and classify utterances in isolation.

We used Google Translate to derive English translations of the Spanish SpCH utterances, and to derive Spanish translations of the English Swbd and MRDA utterances. Of course, translations are far from perfect; DA classification performance could likely be improved by using a translation system trained on spoken dialog. For instance, Google Translate often failed on certain words like “i” that are usually capitalized in text. Even so, when training and testing on translated utterances, the results with the generic system are surprisingly good.

The results reported below used the standard train/test splits provided with the corpora: MRDA had 51 train meetings/11 test; Swbd had 1115 train conversations/19 test; SpCH had 80 train conversations/20 test. The SpCH train set is the smallest at 29k utterances. To avoid issues of differing train set size when comparing performance of different models, we reduced the Swbd and MRDA train sets to the same size as SpCH using randomly selected examples from the full train sets. For each adaptation experiment, we used the target domain training set as the unlabeled data, and report performance on the target domain test set. The test sets contain 4525, 15180, and 3715 utterances for Swbd, MRDA, and SpCH respectively.

4 Results

Table 1 shows the class proportions in the training sets for each domain. MRDA has fewer Backchannels than the the others, which is expected since the meetings are face-to-face. SpCH has fewer Incompletes and more Questions than the others; the reasons for this are unclear. Backchannels have the shortest mean length (less than 2 words) in all domains. Incompletes are also short, while Statements have the longest mean length. The mean lengths of Statements and Questions are similar in the English corpora, but are shorter in SpCH. (This may point to differences in how the utterances were segmented; for instance Swbd utterances can span multiple turns, although 90% are only one turn long.)

Because of the high class skew, we consider two different schemes for training the classifiers, and report different performance measures for each. To optimize overall accuracy, we use basic unweighted training. To optimize average per-class recall (weighted equally across all classes), we use weighted training, where each training example is weighted inversely to its class proportion. We optimize the regularization parameter using a source domain development set corresponding to each training set. Since the optimum values are close for all three domains, we choose a single value for all the accuracy classifiers and a single value for all the per-class recall classifiers. (Different values are chosen for different feature types corresponding to the different adaptation methods.)

	Inc.	**Stat.**	**Quest.**	**Back.**
Swbd	8.1%	67.1%	5.8%	19.1%
MRDA	10.7%	67.9%	7.5%	14.0%
SpCH	5.7%	60.6%	12.1%	21.7%

Table 1: Proportion of utterances in each DA category (Incomplete, Statement, Question, Backchannel) in each domain’s training set.

Table 2 gives baseline performance for all train-test pairs, using translated versions of the test set when the train set differs in language. It also lists the in-domain results using translated (train and test) data, and results using the adaptation methods (which we discuss below). Figure 1 shows details of the contribution of each class to the average per-class recall; bar height corresponds to the second column in Table 2.

4.1 Baseline performance and analysis

We observe first that translation does not have a large effect on in-domain performance; degradation occurs primarily in Incompletes and Questions, which depend most on word order and therefore might be most sensitive to ordering differences in the translations. We conclude that it is possible to perform well on the translated test sets when the training data is well matched. However, cross-domain performance degradation is much worse between pairs that differ in language than between the two English corpora.

We now describe three kinds of issues contributing to cross-domain domain degradation, which we observed anecdotally. First, some highly important words in one domain are sometimes missing entirely from another domain. This issue appears to have a dramatic effect on Backchannel detection across languages: when optimizing for average per-class recall, the English-trained classifiers detect about 20% of the Spanish translated Backchannels and the Spanish classifier detects a little over half of the English ones, while they each detect more than 80% in their own domain.

train set	Acc (%)	Avg. Rec. (%)
Test on Swbd		
Swbd	89.2	84.9
Swbd translated	86.7	80.4
MRDA baseline	**86.4**	**78.0**
MRDA shared only	85.7*	77.7
MRDA SCL	81.8*	69.6
MRDA length only	78.3*	51.4
SpCH baseline	74.5	57.2
SpCH shared only	77.4*	64.2
SpCH SCL	76.8*	**64.8**
SpCH length only	**77.7***	48.2
majority	67.7	25.0
Test on MRDA		
MRDA	83.8	80.5
MRDA translated	80.5	74.7
Swbd baseline	**81.0**	71.6
Swbd shared only	80.1*	**72.1**
Swbd SCL	75.6*	68.1
Swbd length only	68.6*	44.9
SpCH baseline	66.9	50.5
SpCH shared only	66.8	52.1
SpCH SCL	66.1*	**58.4**
SpCH length only	**68.3***	44.6
majority	65.2	25.0
Test on SpCH		
SpCH	83.1	72.8
SpCH translated	82.4	71.3
Swbd baseline	63.8	41.1
Swbd shared only	66.2*	**50.9**
Swbd SCL	68.2*	47.2
Swbd length only	**72.6***	43.6
MRDA baseline	65.1	42.9
MRDA shared only	65.5	**51.2**
MRDA SCL	67.6*	50.9
MRDA length only	**72.6***	44.7
majority	65.3	25.0

Table 2: Overall accuracy and average per-class recall on each test set, using in-domain, in-domain translated, and cross-domain training. Starred results under the accuracy column are significantly different from the corresponding cross-domain baseline under McNemar's test ($p < 0.05$). (Significance is not calculated for the average per-class recall column.) "Majority" classifies everything as Statement.

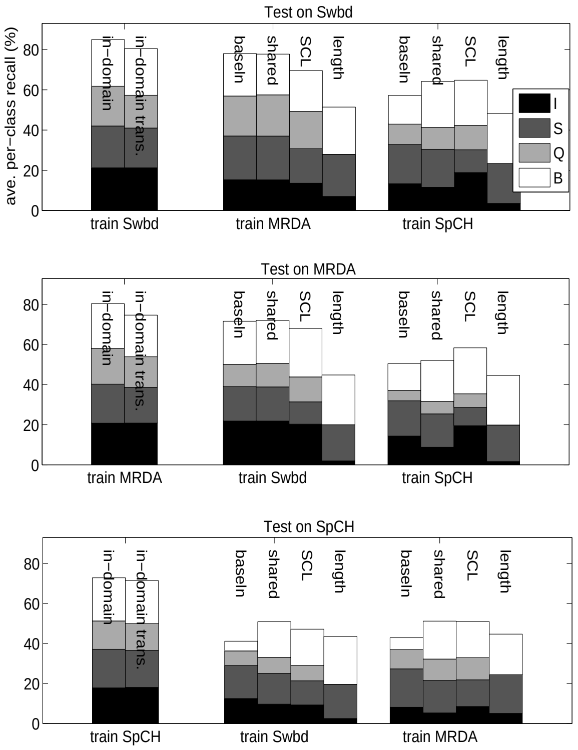

Figure 1: Per-class recall of weighted classifiers in column 2 of Table 2. Bar height represents average per-class recall; colors indicate contribution of each class: I=incomplete, S=statement, Q=question, B=backchannel. (Maximum possible bar height is 100%, each color 25%).

The reason for the cross-domain drop is that many backchannel words in the English corpora (uhhuh, right, yeah) do not overlap with those in the Spanish corpora (mmm, sí, ya) even after translation—for example, "ya" becomes "already", "sí" becomes "yes", "right" becomes "derecho", and "uhhuh", "mmm" are unchanged.

A second issue has to do with different kinds of utterances found in each domain, which sometimes lead to different relationships between features and class label. This is sometimes caused by the translation system; for example, utterances starting with "es que . . ." are usually statements in SpCH, but without capitalization the translator often gives "is that . . .". Since "⟨s⟩–is–that" is a cue feature for Question in English, these utterances are usually labeled as Question by the English domain classifiers. The existence of different types of utterances can result in sets of features that are more highly correlated in one domain than the other. In both Swbd and translated SpCH, utterances containing the trigram "⟨s⟩–but–⟨/s⟩" are most likely to be in the Incomplete class. In Swbd, the bigram "but–⟨/s⟩" rarely occurs outside of that trigram, but in SpCH it sometimes occurs at the end of long (syntactically-incomplete) Statements, so it corresponds to much lower likelihood for the Incomplete class.

The last issue concerns utterances whose true label probabilities given the word sequence are not the same across domains. We distinguish two such kinds utterances. The first are due to class definition differences across domains and annotators, e.g., long statements or questions that are also incomplete are more often labeled Incomplete in SpCH and Swbd than in MRDA. The second kind are utterances whose class labels are not completely determined by their word sequence. To minimize error rate the classifier should label an utterance with its most frequent class, but that may differ across domains. For example, "yes" can be either a Statement of Backchannel; in the English corpora, it is most likely to be a Statement ("yeah" is more commonly used for Backchannels). However, "sí" is most likely to be a Backchannel in SpCH. To measure the effect of differing label probabilities across domains, we trained "domain-general" classifiers using concatenated training sets for each pair of domains. We found that they performed about the same or only slightly worse than domain-specific models, so we conclude that this issue is likely only a minor effect.

4.2 Adaptation using shared features only

In the cross-language domain pairs, some discriminative features in one domain are missing in the other. By removing all features from the source domain training utterances that are not observed (twice) in the target domain training data, we force the classifier to learn only on features that are present in both domains. As seen in Figure 1, this had the effect of improving recall of Backchannels in the four cross-language cases. Backchannels are the second-most frequent class after Statements, and are typically short in all domains. Many typical Backchannel words are domain-specific; by removing them from the source data, we force the classifier to attempt to detect Backchannels based on length alone. The resulting classifier has a better chance of recognizing target domain Backchannels that lack the source-only Backchannel words. At the same time, it mistakes many other short utterances for Backchannels, and does particularly worse on Incompletes, for which length is also strong cue. Although average per-class recall improved in all four cross-language cases, total accuracy only improved significantly in two of those cases, and for the Swbd/MRDA pair, accuracy got significantly worse. The effect on the one-vs.-rest component classifiers was mixed: for some (Statement and some Backchannel classifiers in the cross-language cases), accuracy improved, while in other cases it decreased.

As noted above, the shared feature approach was investigated by Aue and Gamon (2005), who argued that its success depends on the assumption that class/feature relationships be the same across domains. However, we argue here that the success of this method requires stronger assumptions about both the relationship between domains and the correlations between domain-specific and shared features. Consider learning a linear model on either the full source domain feature set or the reduced shared feature set. In general, the coefficients for a given feature will be different in each model—in the reduced case, the coefficients incorporate correlation information and label predictive information for the removed (source-only) features. This is potentially useful on the target domain, provided that there exist analogous, target-only features that have similar correlations with the shared features, and similar predictive coefficients.

For example, consider the discriminative source and target features "uhhuh" and "mmm," which are both are correlated with a shared, noisier, feature (length). Forcing the model to learn only on the shared, noisy feature incorporates correlation information about "uhhuh", which is similar to that of "mmm". Thus, the reduced model is potentially more useful on the target domain, compared to the full source domain model which might not put weight on the noisy feature. On the other hand, the approach is inappropriate in several other scenarios. For one, if the target domain utterances actually represent samples from a subspace of the source domain, the absence of features is informative: the fact that an utterance does not contain "⟨s⟩–verdad–⟨/s⟩", for instance, might mean that it is less likely to be a Question, even if none of the target domain utterances contain this feature.

4.3 Adaptation using SCL

The original formulation of SCL proposed predicting pivot features using the entire feature set, except for those features perfectly correlated with the pivots (e.g., the pivots themselves). Our experiments with this approach found it unsuitable for our task, since even after removing the pivots there are many features which remain highly correlated with the pivots due to overlapping n-grams (i-love vs. love). The number of features that overlap with pivots is large, so removing these would lead to few features being included in the projections. Therefore, we adopted the multi-view learning approach suggested by Blitzer et al. (2009). We split the utterances into two parts; pivot features in the first part were predicted with all the features in the second, and vice versa. We experimented with splitting the utterances in the middle, but found that since the number of words in the first part (nearly) predicts the number in the second part, all of the features in the first part were positively predictive of pivots in the second part so the main dimension learned was length. In the results presented here, the first part consists of the first word only, and the second part is the rest of the utterance. (All utterances in our experiments have at least one word.) Pivot features are selected in each part and predicted using a least-squares linear regression on all features in the other part.

We used the SCL-MI method of Blitzer et al. (2007) to select pivot features, which requires that they be common in both domains and have high mutual information (MI) with the class (according to the source labels.) We selected features that occurred at least 10 times in each domain and were in the top 500 ranked MI features for any of the four classes; this resulted in 78-99 first-part pivots and 787-910 second-part pivots (depending on the source-target pair). We performed SVD on the learned prediction weights for each part separately, and the top (at most) 100 dimensions were used to project utterances on each side.

In all train-test pairs, the first dimension of the first part appeared to distinguish short utterance words from long ones. Such short-utterance words included backchannels from both domains, in addition to acknowledgments, exclamations, swear words and greetings. An analogous dimension existed in the second part, which captured words correlated with short utterances greater than one word (right, really, interesting). The other dimensions of both domains were difficult to interpret.

We experimented with using the SCL features together with the raw features (n-grams and length), as suggested by (Blitzer et al., 2006). As in (Blitzer et al., 2006), we found it necessary to scale up the SCL features to increase their utilization in the presence of the raw features; however, it was difficult to guess the optimal scaling factor without having access to labeled target data. The results here use SCL features only, which also allows us to more clearly investigate the utility of those features and to compare them with the other feature sets.

The most notable effect was an improvement in Backchannel recall, which occurred under both weighted and unweighted training. In addition, there was high confusability between Statements and the other classes, and more false detections of Backchannels. When optimizing for accuracy, SCL led to an improvement in accuracy in three of the four cross-language cases. When optimizing for average per-class recall, it led to improvement in all cross-language cases; however, recall of Statements went down dramatically in all cases. In addition, while there was no clear benefit of the SCL vs. the shared-feature method on the cross-language cases, the SCL approach did much worse than the shared-feature approach on the Swbd/MRDA pair, causing large degradation from the baseline.

As we have noted, utterance length appears to underlie the improvement seen in the cross-language performance for both the SCL and shared-feature approaches. Therefore, we include results for a classifier based only on the length feature. Optimizing for accuracy, this method achieves the highest accuracy of all methods in the cross-language pairs. (It does so by classifying everything as Statement or Backchannel, although with weighted training, as shown in Figure 1, it gets some Incompletes.) However, under weighted class training, the average per-class recall of this method is much worse than the shared-feature and SCL approaches.

Comparison with other SCL tasks Although we basically take a text classification approach to the problem of dialog act tagging, our problem differs in several ways from the sentiment classification task in Blitzer et al. (2007). In particular, utterances are much shorter than documents, and we use position information via the start/end-of-sentence tags. Some important DA cue features (such as the value of the first word) are mutually exclusive rather than correlated. In this way our problem resembles the part-of-speech tagging task

(Blitzer et al., 2006), where the category of each word is predicted using values of the left, right, and current word token. In fact, that work used a kind of multi-view learning for the SCL projection, with three views corresponding to the three word categories. However, our problem essentially uses a mix of bag-of-words and position-based features, which poses a greater challenge since there is no natural multi-view split. The approach described here suffers from the fact that it cannot use all the features available to the baseline classifier—bigrams and trigrams spanning the first and second words are left out. It also suffers from the fact that the first-word pivot feature set is extremely small—a consequence of the small set of first words that occur at least 10 times in the 29k-utterance corpora.

5 Conclusions

We have considered two approaches for domain adaptation for DA tagging, and analyzed their performance for source/target pairs drawn from three different domains. For the English domains, the baseline cross-domain performance was quite good, and both adaptation methods generally led to degradation over the baseline. For the cross-language cases, both methods were effective at improving average per-class recall, and particularly Backchannel recall. SCL led to significant accuracy improvement in three cases, while the shared feature approach did so in two cases. On the other hand, SCL showed poor discrimination between Statements and other classes, and did worse on the same-language pair that had little cross-domain degradation. Both methods work by taking advantage of correlations between shared and domain-specific class-discriminative features. Unfortunately in our task, membership in the rare classes is often cued by features that are mutually exclusive, e.g., the starting n-gram for Questions. Both methods might therefore benefit from additional shared features that are correlated with these n-grams, e.g., sentence-final intonation for Questions. (Indeed, other work on semi-supervised DA tagging has used a richer feature set: Jeong et al. (2009) included parse, part-of-speech, and speaker sequence information, and Venkataraman et al. (2003) used prosodic information, plus a sequence-modeling framework.) From the task perspective, an interesting result is that machine translation appears to preserve most of the dialog-act information, in that in-domain performance is similar on original and translated text.

Acknowledgments

We thank Sham Kakade for suggesting the multi-view SCL method based on utterance splits and for many other helpful discussions, as well as John Blitzer for helpful discussions. We thank the three reviewers for their useful comments.

This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI or the U.S. Government.

References

Anthony Aue and Michael Gamon. 2005. Customizing sentiment classifiers to new domains: a case study. In *Proc. International Conference on Recent Advances in NLP*.

Nuria Bel, Cornelis H. A. Koster, and Marta Villegas. 2003. Cross-lingual text categorization. In *Research and Advanced Technology for Digital Libraries*, pages 126–139. Springer Berlin / Heidelberg.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proc. of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447.

John Blitzer, Dean P. Foster, and Sham M. Kakade. 2009. Zero-shot domain adaptation: A multi-view approach. Technical report, Toyota Technological Institute TTI-TR-2009-1.

Mark G. Core and James F. Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *Proc. of the Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*.

Mark Dredze, John Blitzer, Partha Pratim Talukdar, Kuzman Ganchev, João Graca, and Fernando Pereira. 2007. Frustratingly hard domain adaptation for dependency parsing. In *Proc. of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 1051–1055.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Umit Guz, Gokhan Tur, Dilek Hakkani-Tür, and Sébastien Cuendet. 2010. Cascaded model adaptation for dialog act segmentation and tagging. *Computer Speech & Language*, 24(2):289–306.

Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. 2007. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems 19*, pages 601–608.

Minwoo Jeong, Chin Y. Lin, and Gary G. Lee. 2009. Semi-supervised speech act recognition in emails and forums. In *Proc. of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1250–1259.

Dan Jurafsky, Liz Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical report, University of Colorado at Boulder Technical Report 97-02.

Lori Levin, Ann Thymé-Gobbel, Alon Lavie, Klaus Ries, and Klaus Zechner. 1998. A discourse coding scheme for conversational Spanish. In *Proc. The 5th International Conference on Spoken Language Processing*, pages 2335–2338.

Yang Liu. 2006. Using SVM and error-correcting codes for multiclass dialog act classification in meeting corpus. In *Proc. Interspeech*, pages 1938–1941.

Sinno J. Pan, James T. Kwok, and Qiang Yang. 2008. Transfer learning via dimensionality reduction. In *Proc. of the Twenty-Third AAAI Conference on Artificial Intelligence*.

Brian Roark and Michiel Bacchiani. 2003. Supervised and unsupervised PCFG adaptation to novel domains. In *Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 126–133.

Sophie Rosset, Delphine Tribout, and Lori Lamel. 2008. Multi-level information and automatic dialog act detection in human–human spoken dialogs. *Speech Communication*, 50(1):1–13.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proc. of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26:339–373.

Dinoj Surendran and Gina-Anne Levow. 2006. Dialog act tagging with support vector machines and hidden Markov models. In *Proc. Interspeech*, pages 1950–1953.

Gokhan Tur, Dilek Hakkani-Tür, and Robert E. Schapire. 2005. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45(2):171–186.

Gokhan Tur, Umit Guz, and Dilek Hakkani-Tür. 2007. Model adaptation for dialog act tagging. In *Proc. IEEE Spoken Language Technology Workshop*, pages 94–97.

Gokhan Tur. 2005. Model adaptation for spoken language understanding. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 41–44.

Anand Venkataraman, Luciana Ferrer, Andreas Stolcke, and Elizabeth Shriberg. 2003. Training a prosody-based dialog act tagger from unlabeled data. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume I, pages 272–275.

Nick Webb and Ting Liu. 2008. Investigating the portability of corpus-derived cue phrases for dialogue act classification. In *Proc. of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 977–984.

Dikan Xing, Wenyuan Dai, Gui-Rong Xue, and Yong Yu. 2007. Bridged refinement for transfer learning. In *Knowledge Discovery in Databases: PKDD 2007*, pages 324–335. Springer Berlin / Heidelberg.