Does Active Learning Help Automatic Dialog Act Tagging in Meeting Data?

Anand Venkataraman[1], *Yang Liu*[2], *Elizabeth Shriberg*[1,2], *Andreas Stolcke*[1,2]

[1] Speech Technology and Research Laboratory, SRI International
Menlo Park, CA 94025, USA
[2] International Computer Science Institute, Berkeley, CA 94704, USA
{anand,yangl,ees,stolcke}@speech.sri.com

Abstract

Knowledge of Dialog Acts (DAs) is important for the automatic understanding and summarization of meetings. Current approaches rely on a lot of hand labeled data to train automatic taggers. One approach that has been successful in reducing the amount of training data in other areas of NLP is active learning. We ask if active learning with lexical cues can help for this task and this domain. To better address this question, we explore active learning for two different types of DA models – hidden Markov models (HMMs) and maximum entropy (maxent).

1. Introduction

Annotating conversational dialog act (DA) units for DAs is an errorprone and time-consuming process. Previous work attempted to train automatic DA taggers from substantial amounts of manually tagged data [1, 2, 3] and at striking a compromise between the amount of human-tagged data and the overall accuracy of an automatic tagger using partially supervised training methods [4, 5]. This latter work proposed that the automatic DA tagger would initially be trained (bootstrapped) with a small amount of manually tagged DA data and subsequently refined in an iterative process. Active learning seems to be one of the most appropriate methods to guide the selection of the bootstrap data.

We report on active learning experiments using two different classification paradigms. The first is an extension (and adaptation for active learning) of previous work using hidden Markov model (HMM) based DA taggers. The second is based upon the maximum entropy (maxent) classification principle. In either paradigm, we classify by generating posterior distributions over the DAs. To evaluate classification accuracies we assume that the DA with the highest posterior probability is the selected one. To measure classification uncertainty, we consider the distribution over all the DAs and calculate its entropy. Results are presented of both an exclusively active learning technique and a hybrid technique involving active and partially supervised learning.

2. Previous Work

2.1. Active Learning

Active learning was initially proposed as a way to reduce the number of training examples required to achieve a given degree of performance [6]. It is best described by quoting Cohn et al. [6]: "A learner may proceed by examining the information already given and determining a *region of uncertainty*, an area in the domain where it believes misclassification is still possible. The learner then asks for examples exclusively from that region." A number of applications of this basic technique have surfaced since its introduction, including its application in speech recognition for selecting data to train acoustic phonetic models [7, 8]. In all of the approaches the underlying principle is common – one or more bootstrap models are trained using some minimal set of hand-labeled instances; these bootstrap models are then used to classify a large number of unlabeled instances from which *uncertain classifications* are identified. An oracle (or a human annotator) is then queried for the true labels of these particular instances, which are used to supplement the training data before retraining the classifiers.

2.2. HMM-based DA tagging

The HMM-based DA tagger works by assuming that each session (conversation or meeting) is generated by an HMM in which the various DAs are the states of the HMM. The individual DA units (utterances, u) are considered to be the observations emanating from these states. The likelihoods of the utterances at each state are the probabilities that they can be produced using language models specific to the DA modeled by that state. The transition probabilities are obtained from a language model (which we call the DA grammar) trained from sequences of DAs obtained from the training data. Based on initial experiments on the current corpus, we chose 6-grams for the DA grammar and 3-grams for the DA specific language models and Witten-Bell-smoothing [9] for both.

We adopted the so-called deictic representation of the data, where, in addition to the DAs themselves, the HMM is assumed to also have two special states indicating speaker change or nonchange. Although the speaker nonchange flag is admittedly redundant, we included it for the sake of generality and symmetry. Every utterance is followed by a transition into and out of one of these states. During the decoding process with HMMs, each session is aligned to a specific path within the HMM and the probability, $\mathrm{P}(\mathrm{DA}|u_i)$, that a particular DA state was visited at the time of emission of the DA unit u_i is calculated using the forward-backward algorithm. The entropy of this distribution is indicative of the amount of uncertainty in its classification.

2.3. Maxent DA tagging

The maxent DA tagger works by assuming that each individual utterance, u_i, is characterized by a finite set of features, F_i, which may nor may not include context information (the classification of the previous and next feature set). The posterior probability of the DA tag given these features is estimated us-

Table 1: Data statistics. The number of DA units is averaged over each of the 10 random cuts of data.

Batch size	Number of sessions	Avg. DA Units
Boot	5	7786
Train	50	70316
Validation	10	15149
Test	10	15579

ing the exponential model:

$$\mathrm{P}(DA|F_i) = \frac{e^{\sum_{j=1}^{n} \lambda_j g_j(F_i, DA)}}{Z_i}$$

where the g_j are the indicator functions correponding to the features, the λ_j are the learned feature weights and Z_i is a normalization term.

The maxent model provides an elegant framework to model many correlated features and also capture features that are hard to model using a generative modeling approach. In our maxent classifier, we use the following lexical features: the length of the unit, the identity of the first two words, the identity of the last two words, a bigram of the first two words, the identity of the first word of the next DA unit and a flag indicating whether or not the speaker of the current DA unit is the same as that of the preceding one.

2.4. Data selection for partial supervision

Partially supervised DA tagging as introduced in [4] involves the selection of a small subset of data to manually tag and with which to train bootstrap models. Subsequently, these bootstrap models are used to automatically tag the whole remaining unlabeled data. The automatically obtained tags for the unlabeled DA units are then treated as though they were reference tags and the tagging model is now reestimated based on the entire tagged set. The procedure is iterated a number of times during which the training error typically decreases.

3. Data

We chose data from the ICSI Meeting Corpus, which contains human-annotated dialog act labels for 75 naturally-occurring meetings. Dialog act annotations and associated information are available to the public via the Meeting Recorder Dialog Act (MRDA) corpus [10]. This collection of meetings presents challenges for dialog act modeling, due to its multiple participants, naturalness, high degree of overlap, and different meeting types included. Meetings are roughly an hour in length, and average about 6 participants. DA units were manually classified using a fine grained set of tags and for the purposes of this experiment grouped into 5 broad intuitve classes along the lines of [3] – Backchannels (B), Disruptions (D), Fillers (F), Questions (Q) and Statements (S), with the prior distribution 13.33%, 14.06%, 7.19%, 6.42% and 59.00% respectively.

We divided the 75 annotated sessions into the following classes, each with a specific number of randomly selected sessions as shown in Table 1. The validation set was not used in this set of experiments, but was set aside nevertheless to serve as a basis for continuing experiments on active learning augmented with the partially supervised iterative training procedure discussed in [4].

4. Method

The idea was to train models initially using reference tags from the bootstrap data, and iteratively retrain by supplementing the bootstrap data with up to about 15% of the DA units from the remaining training data. We simulate active learning by ignoring the DA tag information for the training DA units until the reference tags for selected units were specifically requested by the active learner. The bootstrap sessions are critical to the overall tagging procedure because any contextual dialog information we may decide to use is determined solely by them – since the supplemental DA units are chosen on the basis of their entropy, there is no guarantee that any set of subsequently added units will constitute a contiguous sequence in the session that they came from.

The active learning procedure begins by using the bootstrap data set (b DA units). Training the maxent model is straightforward. Features are generated for each of the b units and the maxent parameters are estimated using the EM algorithm from the values of the feature vectors and their desired (target) classification. The DA-grammar for the HMM tagger is likewise trained by extracting sequences of DAs from each of the five bootstrap sessions. The DA specific language models were trained by binning each utterance into its respective DA and then training five DA specific language models, one for each DA. We then use these initial models to generate fast probabilistic classifications for each of the unlabeled units u and calculate the entropy over DAs,

$$H(\mathrm{DA}|u) = \sum_{\mathrm{DA}_i} -\mathrm{P}(\mathrm{DA}_i|u) \log \mathrm{P}(\mathrm{DA}_i|u)$$

where DA_i ranges over the set of all DAs. The DA units are sorted in descending order of entropy and we request reference tags for the top m most entropic units. The training corpus is augmented with these selected units and the classifiers are retrained. The entire process (except the bootstrapping) is repeated n times, at the end of which a total of nm units would have been added to the training set. The final classifier is trained on $b+nm$ DA units, where nm units are actively selected by the learner. We also ran a baseline experiment in which the tagger's DA grammar is trained using the same bootstrap set (b units), but the DA specific language models were trained using and additional nm randomly selected supplemental units. We always ensured that $nm = 10000$, and experimented with procedures that used $m \in \{10, 100, 1000, 10000\}$ units.

In the extension of this procedure to incorporate the partially supervised learning method of [4], we also augment the training data at each iteration with the unselected $T - ni$ units and their hypothesized tags, where T is the total amount of training data and i the iteration number. By uniting the actively selected data with its complement, we create a training set that is complete up to DA sequences, and could thus be used to also reestimate the DA grammar where necessary. In contrast with the approach in [4], we stop iterating once all of the actively selected data has been added into the training corpus. The original approach [4] involved a number of tagging and reestimation steps over the whole data.

Finally, since some stochasticity is involved in all the procedures we have described (the datasets are randomly chosen), each experiment was repeated 10 times using various different *cuts* of the data into bootstrap, training and test sets. The reported numbers are the averaged accuracies from the 10 separate experiments.

4.1. Subtleties specific to the incremental HMM tagger

During the active learning procedure, supplemental units typically move from the unlabeled to the labeled category. It is simple to make this transition with the maxent classifier since their removal from the unlabeled set does not affect the classification of the other units in this set during subsequent iterations. With the incremental HMM tagger, however, this is not possible, since every unit is classified only in the context of its preceding units. We addressed this problem by leaving the units in the unlabeled set despite simultaneously adding them into the labeled set. A further subtlety regarding the HMM tagger is that since supplemental units are picked according to their entropy, although they may be grouped by the sessions they came from, they may not necessarily be in sequence within these sessions. We therefore train the DA grammar only with the bootstrap data and leave it fixed from that point on. Only the DA specific language models are retrained during iteration. In the partially supervised extension to the active learning procedure, however, the DA grammar is also reestimated since the unselected data is also added into the training set with their hypothesized DA tags.

5. Results and Discussion

We found that experiments with 1000 iterations ($n = 1000$) did not provide any significant benefits over and above those with $n = 100$. However, the $n = 100$ experiments did perform significantly better than those with $n = 10$. Since the higher value of n involved a greater number of reestimation steps, we restricted the experiments with the partially supervised extension to only use $n = 100$. Furthermore, the results reported herein are restricted to the set of experiments with $n = 100$ in all cases. The following are the experiments on which we report:

1. A *Ceiling* experiment in which models are trained using the entire available tagged data (approximately 78000 manually tagged units).
2. A *Baseline* experiment in which we randomly pick 10000 manually tagged units to supplement the boot data set.
3. A batch active learning experiment in which 10000 manually tagged units are picked at once on the basis of their high entropy when classified using boot models.
4. An incremental version of the batch active learning experiment, in which 10000 units are added in 100 batches of 100 units each, with model reestimation at each step
5. A partially supervised extension of the incremental active learning experiment in which the actively selected dataset is supplemented with its complement using hypothesized DA tags.

Each of the above experiments was carried out in both the HMM and maxent frameworks. Further, we also report on a version of the incremental active learning experiment in which DA units that repeatedly present with high entropy still count towards m even though they have already been added in a previous iteration.

Table 2 shows the final accuracy at the end of the active learning process with the HMM- and maxent-based tagging procedures. We find that the HMM-based tagger does not benefit from the active learning procedure at all – picking 10000 units at random consistently performs better than either batch or incremental active learning. We discuss possible reasons for this below. In contrast, as Figure 1 shows, we find a significant and consistent gain from the batch active learning procedure with the maxent tagger. Furthermore the gain was modestly amplified by going from batch mode to online processing. We ran experiments to try and isolate or rank the maxent features by usefulness, but found neither redundancies nor severe disparities among the usefulness of the features we had selected.

Table 2: Final tagging accuracies with the maxent and HMM-based taggers after all 10000 DA units have been added into the training data in batches of size n. The baseline accuracies were obtained by supplementing the bootstrap set with 10000 randomly selected utterances. Training data was iteratively supplemented in batches of size $10000/n$ for n iterations.

Batch size	Maxent	HMM
Baseline	78.40	74.65
$n = 1$	78.69	74.41
$n = 10$	78.93	74.42
$n = 100$	79.06	74.47
$n = 1000$	79.04	74.46

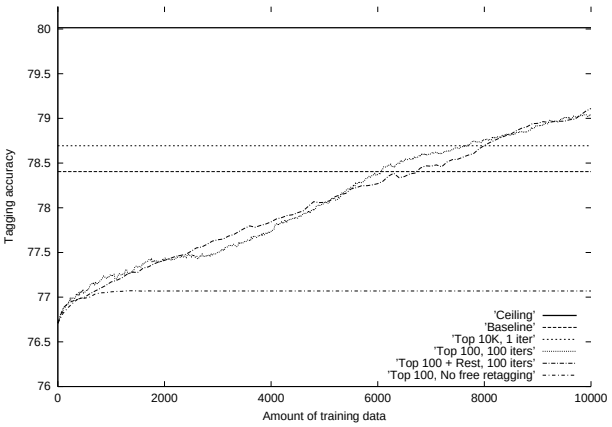

Figure 1: Plot of tagging accuracy with the maxent tagger versus amount of training data.

In Figure 1, the baseline represents the case when the supplemental training data was both randomly chosen and added in a single step. "Top n" indicates that at each iteration, the n most entropic units were added with their reference tags into the training set. Iter indicates the number of iterations of the batch-incremental learning procedure. The partially supervised extension to the active learning procedure is "Top 100 + Rest", which represents the case when the training data is supplemented with both the top 100 most entropic units with reference tags and the rest of the training data with hypothesized tags. Every data point represents an average over 10 independent runs of the experiment to factor out effects stemming from any one particular data division into bootstrap, training and test sets. Our findings indicate that the maxent tagging procedure benefits from both active data selection and incremental model training, but not from the partially supervised extension to active learning. In the *no free retagging* version of our experiment, a supple-

Table 3: Confusion matrix for *recalcitrant* DA units that are repeatedly present with high DA entropy even after being included in the training data. Probabilities are the distributions over manual annotations from the entire hand-tagged training data set. For comparison, a *low* entropy is typically less than about 1e-25. For reference, B = Backchannel, D = Disruption, S = Statement, F = Filler and Q = Question. H() is the entropy of the distribution over the DAs.

Unit	P(B)	P(D)	P(S)	P(F)	P(Q)	$H()$
ah	0.18	0.10	0.68	0.54	0	0.95
but	0	0.39	0	0.61	0	0.69
i mean	0	0.63	0	0.37	0	0.66
mm-hmm	0.88	0.04	0.08	0	0	0.46
no	0.01	0.03	0.85	0.02	0.09	0.59
ok	0.18	0.03	0.67	0.07	0.05	1.01
right	0.39	0.01	0.30	0.03	0.28	1.24
well	0.02	0.44	0.01	0.53	0	0.82
yeah	0.56	0.03	0.34	0.07	0.01	1.02
you know	0.01	0.07	0.22	0.26	0.45	1.26

mental unit that is added into the training set is also left behind in the unlabeled set and could potentially be repicked at later iterations if it presents repeatedly with high entropy. We found that this procedure typically converged prematurely (after 6 to 10 iterations). This happens when the training set consists of at least m utterances that are resistant to further retraining. We were especially intrigued by this observation since one would normally expect at least those units that are part of the training data to be classified with high certainty. We thus repeated the experiment, this time taking care to preserve the *recalcitrant* units. Closer examination of these units showed them to be almost exclusively single word utterances that were intrinsically ambiguous in the absence of prosodic information. Typical examples of such utterances are "right" and "yeah", which, in isolation, could be construed as either answers or questions. In Table 3, we list the confusion matrix for these *recalcitrant* units over the whole human-tagged data. Surprisingly, we found that classification of these units was exceedingly hard even in the presence of automatically determined lexical DA context. In cases like these, we believe that the addition of prosodic cues [5, 11] would help greatly.

Our findings suggest that partially supervised adaptation in addition to the active incremental learning procedure does not contribute significantly to either the HMM- or maxent-based technique. We suspect that the reason for this is twofold: Previous work [4] in partially supervised learning for DA tagging had reported significant gains from iterating, but we did not iterate over and above what was required to add just the actively selected data. Second, [4] also claimed that allowing boot DA tags to drift during the partially supervised learning procedure was beneficial, or indeed that it was detrimental to anchor them to the boot tags. Again, we did not do this, as it would be incompatible with the active learning procedure that added reference units at each step. We surmise from this that the partially supervised learning effort is best decoupled from the active learning phase and allowed to run its course in the normal fashion after the active data selection has completed.

6. Summary

We have described a framework for implementing active data selection for training automatic DA taggers. We also presented a way to couple active learning with partially supervised learning and found that the benefits of this coupling were marginal at best. We reported on experiments using both maxent- and HMM-based DA taggers in the active learning framework and found modest gains with the maxent tagger and no gains with the HMM tagger. Interestingly part of the problem seems to be related to *recalcitrant* DAs, which are inherently ambiguous from text alone, even in the presence of context info. We conclude by suggesting that we may stand to gain by focusing our attention on these inherently ambiguous units and studying ways to effectively classify them, for example using prosodic and semantic cues in addition to the information already exploited.

7. Acknowledgments

We thank Jeremy Ang for help with the MRDA database and features. This work, including infrastructure, was supported by DARPA Contract NBCHD030010, NSF Awards IIS-0121396 and IRI-9619921, the Swiss National Science Foundation through the research network IM2, and by the EU Framework 6 project on Augmented Multi-party Interaction.

8. References

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

[2] G. Ji and J. Bilmes, "Dialog act tagging using graphical models," in *Proc. of ICASSP*, (Philadelpha, PA), March, 2005.

[3] A. Clark and A. Popescu-Belis, "Multi-level dialogue act tags," in *Proc. 5th SIGDIAL Workshop on Discourse and Dialog*, (Boston, MA), 2004.

[4] A. Venkataraman, A. Stolcke, and L. Shriberg, "Automatic dialog act tagging with minimal supervision," in *Proc. 9th Australian International Conference on Speech Science and Technology*, Australian Speech Science and Technology Association, Dec. 2002.

[5] A. Venkataraman, L. Ferrer, A. Stolcke, and E. Shriberg, "Training a prosody based dialog act tagger from unlabeled data," in *Proc. of ICASSP*, 2003.

[6] D. A. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine Learning*, vol. 15, no. 2, pp. 201–221, 1994.

[7] G. Tur, D. Tur, and R. E. Schapire, "Combining active and semi-supervised learning for spoken language understanding," *Speech Communication*, vol. 45, pp. 171–186, 2005.

[8] G. Riccardi and D. Tur, "Active and unsupervised learning for automatic speech recognition," in *Proc. Eurospeech*, 2003.

[9] I. H. Witten and T. C. Bell, "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression," *IEEE Trans. Inf. Theory*, vol. 37, no. 4, pp. 1085–1091, 1991.

[10] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act MRDA corpus," in *Proc. 5th SIGdial Workshop on Discourse and Dialogue*, pp. 97–100, 2004.

[11] E. Shriberg, R. Bates, P. Taylor, A. Stolcke, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. V. Ess-Dykema, "Can prosody aid the automatic classification of dialog acts in conversational speech," *Language and Speech*, vol. 34, no. 3–4, pp. 439–487, 1998.