AUTOMATIC DIALOG ACT LABELING WITH MINIMAL SUPERVISION

Anand Venkataraman Andreas Stolcke Elizabeth Shriberg

Speech Technology and Research Laboratory, SRI International
333 Ravenswood Ave, Menlo Park, CA 94025
E-mail: {anand,stolcke,ees}@speech.sri.com

ABSTRACT: For many natural language applications it is desirable to be able to automatically tag utterances according to their discourse function (dialog act), such as statement, question or acknowledgment. We investigate the problem of automatically tagging dialog acts when hand-labeled training data is scarce. The tagging paradigm employed is a hidden Markov model in which dialog acts are states and utterances are observations, with N-gram language models as observation models. We show that bootstrapping from a small hand-labeled training set, combined with iterative relabeling of a larger unlabeled data set, is an effective approach for preserving accuracy under conditions of limited hand-labeled training data. The dialog act grammar that models the sequencing of dialog acts is found to be of paramount importance in this approach. We analyze the effect that lack of training data has on different dialog act types, and discuss implications for efficient data annotation.

1. INTRODUCTION

For a number of applications in speech recognition and understanding, it is desirable to be able to classify utterances according to their discourse function, into so-called *dialog acts* (DAs), such as statements, questions, or acknowledgments. There is a long history of applying various statistical modeling and machine learning approaches for automatic DA tagging (see Stolcke et al. (2000) for a review), but all of these approaches have in common their requirement for significant hand-labeled training data. Obtaining such data can be time-consuming and expensive, especially if it involves listening to the speech data. This motivates us to look for ways to use *unlabeled* data for the training of DA tagging models, that is, to train DA taggers with *minimal supervision*. Note that completely unsupervised methods are also of interest; however, we can assume that a minimal set of hand-labeled data will always be required if the goal is to tag a predefined (e.g., a meaningful or interpretable) set of DA classes.

In this paper, we describe an algorithm for minimally supervised DA tagging that is based on hidden Markov models (HMMs), N-gram DA language models, and iterative reestimation of unlabeled data. We study its performance on a spoken dialog corpus. Section 2 reviews previous work, much of which we build upon. Section 3 describes our corpus and data. Section 4 describes the model and the iterative tagging procedure. Section 5 discusses results, and Section 6 provides a summary and suggestions for future work.

2. PREVIOUS WORK

The HMM-based approach to DA tagging with N-gram language models goes back to at least Nagata (1992) and Nagata and Morimoto (1993), with many variations since (see Stolcke et al. (2000)). HMM tagging models have also been used successfully for part-of-speech labeling (Church 1988), where a priori restrictions on the possible labels (POS tags) of observations (words) can be leveraged for unsupervised training (Cutting, Kupiec, Pedersen, and Sibun 1992). However, such a weakly supervised approach is not applicable to DA tagging since in the latter case the observations are word sequences with a great deal of variability, so that it is not usually possible to define a priori restrictions on possible DA tags given the words.

Lobacheva (2000) has investigated automatic reclustering of word sequences starting with hand-labeled DA classes, in order to improve the DA-models in terms of their language model perplexity and speech recognition accuracy. However, in that work the goal was not DA tagging accuracy per se, and no dialog act grammar (DA sequence model) was employed in the reclustering.

Proceedings of the 9th Australian International Conference on Speech Science & Technology
Melbourne, December 2 to 5, 2002.

Table 1: List of DAs and their distribution. Vocabulary size is the total number of distinct words within that class in the training data.

DA	Description	Proportion	Vocab. size
SS	Action statements	23.55%	211
SF	Target location statements	15.86%	183
SH	Status reports	10.74%	74
SO	All other statements	7.12%	304
Q	Questions/Requests for Acknowledgment	7.98%	190
A	Answers and Acknowledgments	34.75%	86

3. DATA

Data for our experiments was drawn from the SPINE (Speech in Noisy Environments) corpus, which was created for developing speech recognition in military noise environments (Navy Research Laboratory 2001). The data consists of spoken dialogs between two collaborating humans in a simulated battle game. The most recent development and evaluation test sets for the SPINE task were set aside for eventual recognition experiments, and were not used in this study. The rest of the corpus consisted of 324 sessions (dialogs) containing 29,047 utterances (sessions averaged about 90 utterances each).

A SPINE dialog typically consists of one or more questions, commands, or requests for clarification by one party, followed by answers, status reports, or acknowledgments by the other. However, many of the utterances in this data set were found to consist of what we would intuitively label as more than one dialog act. For example, the utterance

um. yeah. i'll i'll sweep it now just just to uh check.

consists of three dialog acts: a hesitation (UM), a YES-NO answer, and a statement about an action to be performed. We therefore segmented the data into dialog act units, which can be done reasonably well automatically by rules based on punctuation and key words such as “Um” and “Yeah”. We thus obtained 44,412 DA tokens in the 324 dialogs. Henceforth, we use the words “utterances” and “sessions” synonymously with the resegmented units and their dialogs, respectively.

A small subset of this corpus was hand labeled for DA classes for the purposes of this study. This hand labeled corpus consisted of randomly selected sets of 20 contiguous original utterances from 89 randomly selected sessions. We thus obtained a set of 2,794 dialog acts, which were annotated by one of the authors using a detailed set of discourse tags. Note that while the DA models described later are based only on the word transcripts associated with the data, the hand labeling was based on listening to the utterances, so as to better disambiguate their discourse functions. The detailed DA annotations were subsequently mapped into a smaller set of six broad DA classes, on which all further work was based. Table 1 shows the resultant DA classes and their distribution in the hand-labeled data.

Five different subsets of the hand-labeled data were created to study the effect of reducing the amount of available bootstrap data. To keep the results comparable, however, the same test set was used in all experiments regardless of the amount of bootstrap data. The test set consisted of 1,192 hand-labeled utterances (37 sessions). Table 2 shows the boot data sizes for the various experiments. Finally, we divided the remaining unlabeled 32,745 DAs (235 sessions) into a 29,471 DA (209 sessions) set for unsupervised training, and a 3,274 DA (26 sessions) set, for cross-validation during training, as described later.

4. MODEL AND METHOD

HMM tagging applies naturally to DA tagging; we outline the framework here, and refer to the cited literature for details. The DAs are represented by the model's hidden states, whereas the utterances

Table 2: List of various amounts of bootstrap data tested.

Expt. name	**Boot utterances**	**Boot sessions**	**% of Training**
Boot-1602	1602	52	5.39%
Boot-242	242	8	0.81%
Boot-146	146	4	0.49%
Boot-72	72	2	0.24%
Boot-44	44	1	0.15%

in the DAs correspond to the observations generated by the states. To model the interaction between speakers, the state space is augmented to also encode which of the participants is speaking. Two types of N-gram language models (LMs) are employed. For each DA class, the observation probabilities (of words given the DA type) are estimated by an N-gram model trained on words from that DA class. Second, the transition probabilities between states (DAs) are estimated by an N-gram model over DA labels, the *dialog grammar*. Thus, a trigram dialog grammar would predict each DA type from the two preceding DA types. A unigram dialog grammar uses only prior probabilities for the DA types, and does not make use of dialog context.

The standard supervised approach trains the model components (DA N-grams and DA grammar) from labeled data, and tags the test data by decoding the most likely DA sequence for the given string of observed wordss (using the Viterbi algorithm). A partially supervised version that uses additional, unlabeled training data is also straightforward. In that approach, an initial tagger is trained from the available labeled data (the bootstrap data); it is then used to tag the unlabeled training data, after which the model is retrained using all training data. The resulting tagger is presumably improved over the initial one that was trained on the bootstrap data alone; we should therefore iterate the training and relabeling.

Evaluation metrics

Instead of viewing the repeated retraining procedure as an iterative improvement of tagger accuracy, we can also see it as an approximate form of expectation-maximization (EM) (Dempster, Laird, and Rubin 1977) aimed at maximizing the likelihood of the training data. The approximation uses the single most probable hidden states (DAs) in the maximization step, instead of their posterior distribution. We therefore reason that the likelihood of *unlabeled* held-out data, as computed by the HMM forward dynamic programming algorithm, can be used to cross-validate the model at each iteration—preventing overfitting to the training data. This is fortunate because *labeled* held-out data is typically a precious resource and may be better utilized for bootstrapping.

On the independent test data we evaluate models in terms of both tagging accuracy (relative to hand labeling) and data likelihood. Likelihoods are expressed more intuitively as *perplexities*, or average word branching factors:

$$\mathrm{ppl(D)} = \mathrm{e}^{-\ln \mathrm{P(D)}/|\mathrm{D}|} \tag{1}$$

where $\mathrm{P}(D)$ is the probability of the data under the model, summing over all possible DA sequences. Note that perplexity evaluates the DA models at the word level, as a language model, without regard to DA labeling accuracy.

5. RESULTS AND DISCUSSION

We evaluated the model and training procedure for different amounts of bootstrap data, and for two different dialog grammars. Both unigram (no DA context) and trigram (context is the two previous DAs) dialog grammars were used to assess the importance of dialog context for this task. No benefit was found from using higher-order dialog grammars. The DA models were implemented as word trigram models throughout. All N-gram models were smoothed with backoff (Katz 1987) and Witten-Bell discounting (Witten and Bell 1991). For bootstrap data, the two extreme conditions were *Boot-1602*, with 1602 labeled utterances for initial training, and *Boot-44*, with a mere 44 utterances.

The accuracies as a function of training iteration for the four combinations are plotted in Figure 1.

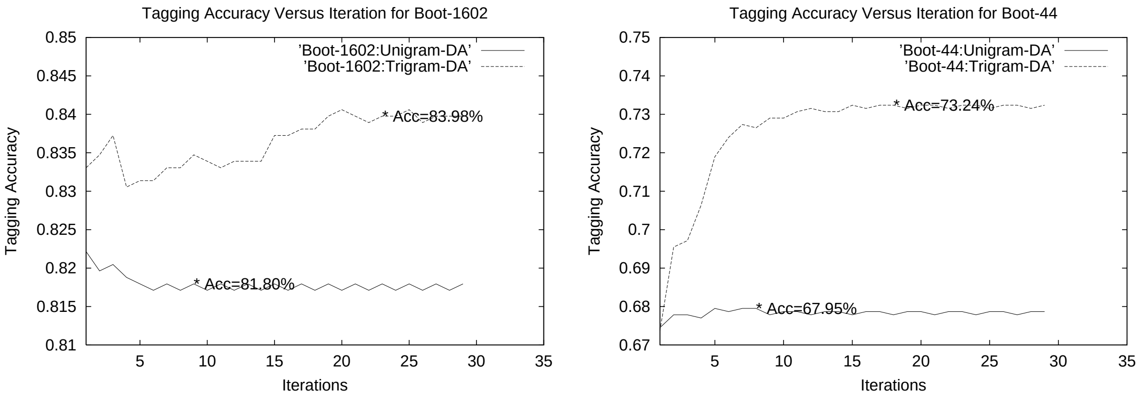

Figure 1: Accuracy versus iteration for the Boot-1602 and Boot-44 experiments, respectively. Asterisks indicate the points at which validation set perplexity was minimized, and the accompanying captions give the test set accuracy at those iterations. The accuracies after iteration 1 in the two cases were 82.6% and 83.1% for Boot-1602:Unigram-DA and Boot-1602:Trigram-DA, and 66.6% and 66.2% for Boot-44:Unigram-DA and Boot-44:Trigram-DA, respectively.

In Table 3, we compare the evaluation metrics (computed on the test set) after the first training iteration (which uses only the bootstrap data) to that of the best iteration as determined by cross-validation. We show here the results for all the five different amounts of bootstrap data.

Table 3: Summary of results for unigram and trigram HMMs taggers, and for different amounts of bootstrap data. "Iter" shows the iteration at which the best validation set perplexity was found. Accuracies and perplexities are given for the classification using the model selected at that iteration, as well as after iteration 1.

	1-gram dialog grammar					**3-gram dialog grammar**				
EXPT name	**Iter**	Accuracy		Perplexity		**Iter**	Accuracy		Perplexity	
		Best	Iter-1	Best	Iter-1		Best	Iter-1	Best	Iter-1
Boot-1602	3	82.05	82.63	16.72	17.02	19	83.98	83.05	15.91	16.21
Boot-242	7	75.67	75.17	16.62	17.10	13	79.03	73.32	15.97	16.72
Boot-146	5	67.36	67.03	16.76	17.14	26	71.22	65.52	15.69	16.66
Boot-72	6	71.89	70.64	17.09	17.43	15	73.57	69.88	16.60	17.17
Boot-44	5	67.95	66.61	17.08	17.28	13	73.07	66.19	16.65	17.07

We draw a number of observations from these results. First, we note from the marked points on the accuracy plots (Figure 1) that our cross-validation criterion (likelihood of unlabeled data) does a good job of picking models with near-optimal classification accuracy on the test data, in all cases where iterating is helpful.

Second, iterative retraining always seems to help, without overfitting, except in the condition with the largest amount of hand-labeled training data and when no dialog context is used (unigram dialog grammar). The latter case seems to indicate that when the initial tagger performs relatively well, but tagging accuracy on the unlabeled data is not high enough (e.g., due to lack of discourse context), then the addition of the unlabeled data can actually hurt performance.

Third, accuracy improvements due to retraining increase as less hand-labeled data is used. However, using only the full amount of available hand-labeled data (Boot-1602, iteration 1) is always better than using less bootstrap data and a much larger amount of unlabeled training data. In other words: "There is no data like labeled data".[1]

Also, we observe that paucity of training data adversely affects the unigram-DA tagger much more seriously than the trigram-DA tagger. Coupled with the fact that the trigram version is only slightly better with the full hand-labeled training set (Boot-1602), we can interpret this as follows: the DA-specific N-gram models are a strong knowledge source compared to the dialog act grammar, but they require considerable training data to be effective. The DA grammar, which has a much smaller vocabulary and fewer parameters, can be effective even with minimal training data. It can thus partly compensate for the overall lack of training data, especially when combined with the unlabeled data.

In these experiments we allowed the retraining procedure to relabel all training data, including the initially hand-labeled bootstrap set. Surprisingly, this yielded better results—even on accuracy—than when the initial labels were fixed. This could indicate that our hand labels may be somewhat noisy (inconsistent), and that relabeling them produces sharper models.

In Table 4 we show the effect of diminishing bootstrap data on classification accuracies per DA, for the trigram dialog grammar. As the amount of bootstrap data decreases from 1602 to 44 utterances, we see a 10.91% absolute degradation in classification accuracy. Interestingly, this difference is not evenly shared among the DA classes. In the last row of Table 4, we weight the contribution to total degradation due to each DA by their proportions in the training data. As shown, the SH, SF and A classes contribute much less to the total degradation than do the SO, Q and SS classes.[2] In fact, we even see a marginal improvement in the classification of SF with the reduced bootstrap set size. This result, however, is not totally unexpected. The proportions in the training data of both SO and Q are small. By reducing the size of the bootstrap data, we reduce even more significantly the numbers of the already modestly represented classes within it. Indeed, there is only one instance of Q and one instance of SO in the reduced set, which makes learning anything about these classes clearly not feasible. A question, however, is why SS (which is relatively better represented in the reduced set, with 13 instances) also contributes to decreased accuracy. We suspect that this is due to a subtle interplay between amounts of reductions in the data and the relative perplexities within each of the classes. If the data perplexity within a class is high, one would expect to require a correspondingly greater amount of training data to be able to successfully generalize a model over that class. If further investigation yields any insight into this phenomenon, it could form the basis of very useful guidelines for data annotation when annotation resources are limited.

Table 4: Contributions to total performance degradation by DA type when available bootstrap data is reduced. Weighted difference is the absolute difference in accuracy, weighted by the proportion of the DA in the full hand-labeled set of 2794 utterances.

EXPT name	**Total**	**SH**	**A**	**SO**	**Q**	**SF**	**SS**
Boot-1602	83.98	92.36	93.72	78.89	39.51	85.16	80.51
Boot-44	73.07	86.81	90.84	0.00	0.00	87.36	76.68
Percent of Training	100.00	10.74	34.75	7.12	7.98	15.86	23.55
Absolute difference	10.91	5.55	2.88	78.89	39.51	-2.20	3.83
Weighted difference	10.91	0.60	1.00	5.62	3.15	-0.35	2.97

[1] There is an anomaly in the Boot-146 results, which are worse than those with less boot data. Upon inspection of the dialogs involved we found that this was due to an unfortunate random selection of the Boot-146 set, in which two of the four dialogs have an unusually high proportion of the "SO" class. (The preponderance of SO tags in these sessions occurred because the battle simulator was not working; participants were thus trying to figure out what was wrong, instead of actually playing the game.) This results in a severe mismatch between the training and test sets. The Boot-72 and Boot-44 sets do not contain these anomalous dialogs and are therefore spared their sabotaging consequences.

[2] Summing the DA specific accuracies weighted by their proportions as shown in Table 4 will not necessarily equal the overall accuracy. The reason for this is that the proportions were obtained from the full hand-labeled set (as shown in Table 1), while the accuracies are computed on the evaluation set only. The proportions of these sets, though similar, are not identical.

6. CONCLUSIONS AND FUTURE WORK

We have investigated an approach for classifying spontaneously spoken utterances into dialog act categories when large amounts of unlabeled data are available, but only a small amount of hand-labeled training data is available. The algorithm is based on iterative relabeling of the training data, and model reestimation. We have shown the effectiveness of the approach on a corpus of task-oriented spoken dialogs. The approach is especially effective when combined with a dialog act grammar that uses context in DA tagging. We also find that relatively small amounts of labeled data can be more effective than much larger amounts of unlabeled data. In ongoing work, we are exploring the addition of further knowledge sources to the tagger in a partially supervised paradigm. In particular, we aim to integrate prosodic cues, such as those used in (Shriberg et al. 1998), which have proven useful in fully supervised DA modeling. We are also investigating strategies for optimizing the labeling effort, so that hand-annotated data can be maximally effective in classifier training.

References

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*(1), 1–38.

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing ASSP-35*(3), 400–401.

Lobacheva, Y. (2000). *Discourse Mixture Language Modeling*. Unpublished Masters dissertation, Boston University, Boston, MA.

Navy Research Laboratory (2001). Speech in Noisy Environments. http://elazar.itd.nrl.navy.mil/spine/.

Shriberg, E., R. Bates, P. Taylor, A. Stolcke, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. V. Ess-Dykema (1998). Can prosody aid the automatic classification of dialog acts in conversational speech. *Language and Speech 34*(3–4), 439–487.

Stolcke, A., N. Coccaro, R. Bates, P. Taylor, C. V. Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer (2000). Dialogue act modeling for autoamtic tagging and recognition of conversational speech. *Computational Linguistics 26*(3), 339–373.

Witten, I. H. and T. C. Bell (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory 37*(4), 1085–1091.