

EXPLOITING UNLABELED UTTERANCES FOR SPOKEN LANGUAGE UNDERSTANDING

Gokhan Tur Dilek Hakkani-Tür

AT&T Labs – Research
180 Park Avenue, Florham Park, NJ 07932 USA
{gtur,dtur}@research.att.com

ABSTRACT

State of the art spoken language understanding systems are trained using labeled utterances, which is labor intensive and time consuming to prepare. In this paper, we propose methods for exploiting the unlabeled data in a statistical call classification system within a natural language dialog system. The basic assumption is that some amount of labeled data and relatively larger chunks of unlabeled data is available. The first method augments the training data by using the machine-labeled call-types for the unlabeled utterances. The second method, instead, augments the classification model trained using the human-labeled utterances with the machine-labeled ones in a weighted manner. We have evaluated these methods using a call classification system used for AT&T natural dialog customer care system. For call classification, we have used a boosting algorithm. Our results indicate that it is possible to obtain the same classification performance by using 30% less labeled data when the unlabeled data is utilized. This corresponds to a 1-1.5% absolute classification error rate reduction, using the same amount of labeled data.

1. INTRODUCTION

Spoken dialog systems aim to identify intents of humans, expressed in natural language, and take actions accordingly, to satisfy their request. In a natural spoken dialog system, first the speaker's utterance is recognized using an automatic speech recognizer. Then, the intent of the speaker is identified from the recognized sequence, using a natural language understanding component. This step can be seen as a call routing or a classification problem [1]. In this study, we have used a boosting-style classification algorithm [2]. As a call classification example, consider the utterance *I would like to learn my account balance*, in a customer care application. Assuming that the utterance is recognized correctly, the corresponding intent or the call-type would be *Account Balance Request* and the action would be prompting the balance to the user or routing this call to the billing department.

When statistical classifiers are used in such systems (e.g., [3, 4]), they are trained using large amounts of task data

which is usually transcribed and then labeled by humans, a very expensive and laborious process. By "labeling", we mean assigning one or more of the predefined call-type(s) to each utterance. It is clear that the bottleneck in building an accurate statistical system is the time spent for high quality labeling.

Building better call classification systems in a shorter time frame motivates us to develop novel techniques. In this paper, we present two semi-supervised learning methods for combining labeled and unlabeled utterances, for speeding up the building of accurate call-type classification systems. The first method simply adds the machine labeled utterances to the training data. The second method is specific to the boosting algorithms and augments the classification model trained using the human-labeled utterances with the machine-labeled ones in a weighted manner.

In the following section, we summarize the previous approaches combining labeled and unlabeled data for related machine learning problems and review some of the related work in language processing. In Section 3, we briefly explain boosting algorithms. Then, in Section 4, we present the methods we propose. We conclude with our experiments, results, and a discussion of some of the future work.

2. RELATED WORK

Recently semi-supervised learning algorithms that use both labeled and unlabeled data have been used for text classification in order to reduce the need for labeled training data. Blum and Mitchell [5] have used the Co-Training approach for web page classification to boost the performance of the learning algorithm when only a small number of examples are available. For using Co-Training, the features in the problem domain should naturally divide into two sets. For the same task, Nigam *et. al.* [6] have used an algorithm for learning from labeled and unlabeled documents based on the combination of Expectation Maximization (EM) and Naive Bayes classifier. Ghani [7] has combined the EM algorithm as well as Co-Training with Error-Correcting Output Coding, to exploit the unlabeled data, in addition to the labeled data.

For natural language call routing, Iyer *et. al.* has proposed using speech recognizer output instead of transcribing the utterances during training, without losing accuracy. However, hand-labeling of the utterances with the correct call-type is not mentioned.

Another approach to reducing the amount of labeled data for classification is active learning, where the aim is to select the most informative examples for classification performance improvement prior to labeling and label only them [8]. Our previous work includes using certainty-based active learning approaches for reducing the amount of labeled data needed for spoken language understanding [9] and automatic speech recognition [10].

Combining the data with prior task knowledge (e.g. rules) is also considered in the literature for building natural language dialog systems in a shorter time frame. Schapire and others have extended boosting so as to handle initial hand-written rules during classification [11]. When there is little labeled data, this approach is shown to be very effective.

3. BOOSTING

We begin by a review of boosting-style algorithms. Boosting aims to combine “weak” base classifiers to come up with a “strong” classifier. This is an iterative algorithm, and in each iteration, a weak classifier is learned so as to minimize the training error.

More formally, the algorithm (for the simplified binary (+1 and -1) classification case) is as follows:

- Given the training data from the instance space X : $(x_1, y_1), \dots, (x_m, y_m)$ where $x_i \in X$ and $y_i \in \{-1, +1\}$
- Initialize the distribution $D_1(i) = 1/m$
- For each iteration $t = 1, \dots, T$ do
 - { Train a base learner, h_t , using distribution D_t .
 - { Update $D_{t+1}(i) = D_t(i)e^{-\alpha_t y_i h_t(x_i)} / Z_t$ where Z_t is a normalization factor and α_t is the weight of the base learner.

- Then the output of the final classifier is defined as:

$$H(x) = \text{sign}(f(x)) \text{ where } f(x) = \sum_{t=1}^T \alpha_t h_t(x)$$

This algorithm can be seen as a procedure for finding a linear combination of of base classifiers which attempts to minimize a loss function, which in this case is:

$$\sum_i e^{-y_i f(x_i)}$$

An alternative would be minimizing logistic loss, which is:

$$\sum_i \ln(1 + e^{-y_i f(x_i)})$$

A more detailed explanation and analysis of this algorithm can be found in [2].

4. APPROACH

In this work, the aim is to exploit the unlabeled utterances in a semi-supervised fashion. To this end, we propose two methods. Both methods assume that there is some amount of training data available for training an initial classifier. The basic idea is to use this classifier to label the unlabeled data automatically, and improve the classifier performance using the machine-labeled call-types as the labels of those unlabeled utterances, thus reduce the amount of human-labeling effort necessary to come up with decent statistical systems.

4.1. Augmenting the Data

This is the simpler method. First we train an initial model using the human-labeled data, and then classify the unlabeled ones. Then we add the unlabeled utterances directly to the training data, by using the machine-labeled call-types as seen in Figure 1. In order to reduce the noise added because of classifier errors, we only add those utterances which are classified with the call-types with a confidence higher than some threshold. This threshold can be set using a separate held-out set. Then whole data including both human- and machine-labeled utterances are used for training the classifier again.

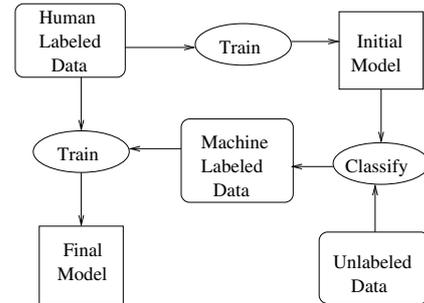


Fig. 1. The first method: Augmenting the data

4.2. Augmenting the Classification Model

For the second method for semi-supervised learning we again train a classifier using a small amount of labeled data. Then, we augment that model by unlabeled examples in a weighted manner. Figure 2 depicts the process proposed for this method.

This method is similar to incorporating prior knowledge into boosting [11]. In that work, a model which fits both the training data and the task knowledge is trained. In our case, the aim is to train a model that fits both the human-labeled and machine-labeled data. For this purpose, we first train an initial model using the human-labeled data. Then, the boosting algorithm tries to fit both the machine-labeled data

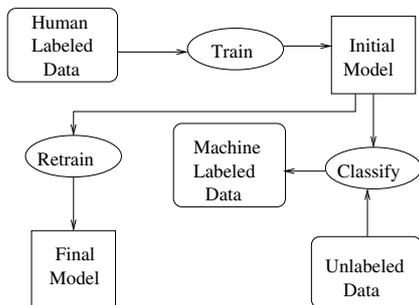


Fig. 2. The second method: Augmenting the model

and the prior model using the following loss function:

$$\sum_i (\ln(1 + e^{-y_i f(x_i)}) + \eta KL(P(\cdot|x_i) \parallel \rho(f(x_i))))$$

where

$$KL(p \parallel q) = p \ln(p/q) + (1 - p) \ln((1 - p)/(1 - q))$$

is the Kullback-Leibler divergence (or binary relative entropy) between two probability distributions p and q . In our case, they correspond to the distribution from the prior model, $P(\cdot|x_i)$, to the distribution from the constructed model, $\rho(f(x_i))$. This term is basically the distance from the initial model built by human-labeled data and the new model built with machine-labeled data. In the marginal case, if these two distributions are always the same then the KL term will be 0 and the loss function will be exactly the same as the first term, which is nothing but the logistic loss. η is used to control the relative importance of these two terms. This weight may be determined empirically on a held-out set. In addition to that, similar to the first method, in order to reduce the noise added because of classifier errors, we can only exploit those utterances which are classified with a confidence higher than some threshold.

5. EXPERIMENTS AND RESULTS

We have evaluated these semi-supervised learning methods using the utterances from the database of the *How May I Help You*SM (*HMIHY*SM) system for AT&T customer care [12]. In this natural dialog system, users are asking questions about their phone bills, calling plans, etc., and the system aims to classify them into one or more of the 49 call-types in total, such as *Account Balance Request*, or *Calling Plans*. There are 57,829 utterances in the training data, 3,500 utterances in the held-out set, and 3,513 utterances in the test set. All of them are transcribed. We have performed our tests using the Boostexter tool [4]. For all experiments, we have used word n -grams as features and iterated 500 times.

First, we have selected the optimal threshold of top scoring call-type confidences using the held-out set. Obviously

there is a trade-off in selecting the threshold. If it is set to a lower value, that means a larger amount of noisy data, and if it is set to a higher value, that means less amount of useful or informative data. Figure 3 proves this behavior for the held-out set. We have trained initial models using 2,000, 4,000, and 8,000 human-labeled utterances and then augmented these as described in the first method, with the remaining data (only using machine-labeled call-types). On the x axis, we have different thresholds to select from the unlabeled data which the classifier uses, and on the y axis we have the classification error rate if that data is also exploited. Classification error rate is the ratio of utterances for which the classifier’s top scoring call-type is not one of the correct labels. A threshold of 0 means using all the machine-labeled data and 1 means using none. As seen, there is consistently 1-1.5% difference in classification error rates using various thresholds for each data size.

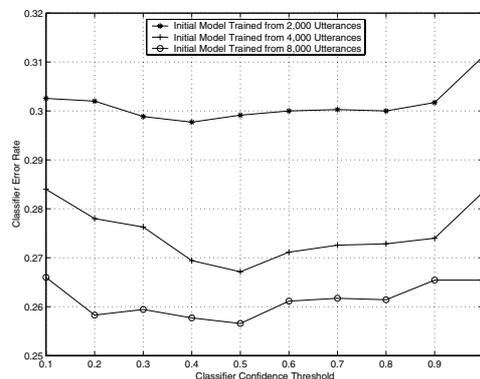


Fig. 3. Trade-off for choosing the threshold to select among the machine-labeled data on the held-out set.

Figure 4 depicts the performance using the two methods proposed, by plotting the learning curves for various initial labeled data set sizes. In the figure, x axis is the amount of human-labeled training utterances, and y axis is the classification error rate of the corresponding model on the test set. The baseline is the top curve, with the highest error rate, where no machine-labeled data is used. The two curves below the baseline are obtained by the two proposed methods. In both methods, we selected 0.5 as the threshold for selecting machine labeled data, and for each data size, we optimized the weight η in the second method using the held-out set. As in case of the held-out set, we have consistently obtained 1-1.5% classifier error rate reductions on the test set using both approaches when the labeled training data size is less than 15,000 utterances. The reduction in the need for human-labeled data to achieve the same classification performance is around 30%. For example we have got the same performance when we have used 5,000 human-labeled utterances instead of 8,000 if we augment the data with unlabeled utterances.

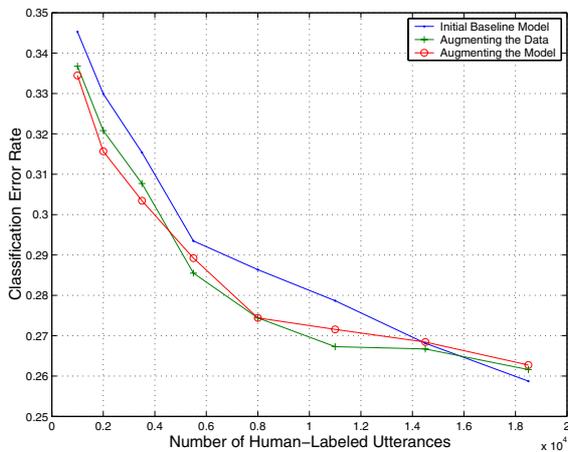


Fig. 4. Results using semi-supervised learning. Top most learning curve is obtained using just human-labeled data as a baseline. Below that lie the learning curves using the first and second methods.

6. ACKNOWLEDGMENTS

We would like to thank Robert E. Schapire for providing us Boostexter classifier and for many helpful discussions.

7. CONCLUSIONS AND DISCUSSION

We have presented methods for exploiting the unlabeled speech utterances, which are confidently classified by the classifier. We have shown that, for the task of call classification, using these semi-supervised learning methods, it is possible to improve the performance of a spoken language understanding system. Note that, most classifiers support a way of combining models or augmenting the existing model, so although this implementation is classifier (boosting) dependent, the idea is more general. Our results indicate that we have achieved the same call classification accuracy using 30% less labeled data when there is not much training data available.

The challenge with semi-supervised learning is that only the utterances which are classified with a confidence larger than some threshold may be exploited in order to reduce the noise introduced by the classifier errors. Intuitively, the noise introduced would be less with better initial models, but in such a case, additional data will be less useful. So one may expect such semi-supervised techniques to work less with very little or very large amounts of data. An alternative approach for using a threshold to select machine-labeled data would be modifying the classifier, so that at each iteration, the confidence of the call-types contribute to the data distribution.

One problem with these approaches is that, a call-type may be poorly trained using the initial human-labeled data,

as there is very little or no data for that call-type. The proposed approaches are not supposed to improve the classification accuracy for such call-types. Our future work includes attacking and exploring those issues.

8. REFERENCES

- [1] G. Tur, J. Wright, A. Gorin, G. Riccardi, and D. Hakkani-Tür, "Improving spoken language understanding using word confusion networks," in *Proceedings of the ICSLP*, Denver, CO, September 2002.
- [2] R. E. Schapire, "The boosting approach to machine learning: An overview," in *Proceedings of the MSRI Workshop on Nonlinear Estimation and Classification*, 2002.
- [3] P. Haffner, G. Tur, and J. Wright, "Optimizing SVMs for complex call classification," in *Proceedings of the ICASSP*, Hong Kong, April 2003.
- [4] 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.
- [5] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the Workshop on Computational Learning Theory (COLT)*, Madison, WI, July 1998.
- [6] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, vol. 39, no. 2/3, pp. 103–134, 2000.
- [7] R. Ghani, "Combining labeled and unlabeled data for multiclass text categorization," in *Proceedings of the ICML*, Sydney, Australia, July 2002.
- [8] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Proceedings of the ICML*, New Brunswick, NJ, July 1994.
- [9] G. Tur, R. E. Schapire, and D. Hakkani-Tür, "Active learning for spoken language understanding," in *Proceedings of the ICASSP*, Hong Kong, May 2003.
- [10] D. Hakkani-Tür, G. Riccardi, and A. Gorin, "Active learning for automatic speech recognition," in *Proceedings of the ICASSP*, Orlando, FL, May 2002.
- [11] R. E. Schapire, M. Rochery, M. Rahim, and N. Gupta, "Incorporating prior knowledge into boosting," *Proceedings of the ICML*, July 2002.
- [12] A. L. Gorin, G. Riccardi, and J. H. Wright, "Automated natural spoken dialog," *IEEE Computer Magazine*, vol. 35, no. 4, pp. 51–56, April 2002.