Semi-supervised Speech Act Recognition in Emails and Forums

Minwoo Jeong[†*] **Chin-Yew Lin**[‡] **Gary Geunbae Lee**[†]
[†]Pohang University of Science & Technology, Pohang, Korea
[‡]Microsoft Research Asia, Beijing, China
[†]{stardust,gblee}@postech.ac.kr [‡]cyl@microsoft.com

Abstract

In this paper, we present a semi-supervised method for automatic speech act recognition in email and forums. The major challenge of this task is due to lack of labeled data in these two genres. Our method leverages labeled data in the Switchboard-DAMSL and the Meeting Recorder Dialog Act database and applies simple domain adaptation techniques over a large amount of unlabeled email and forum data to address this problem. Our method uses automatically extracted features such as phrases and dependency trees, called subtree features, for semi-supervised learning. Empirical results demonstrate that our model is effective in email and forum speech act recognition.

1 Introduction

Email and online forums are important social media. For example, thousands of emails and posts are created daily in online communities, e.g., Usenet newsgroups or the TripAdvisor travel forum[1], in which users interact with each other using emails/posts in complicated ways in discussion threads. To uncover the rich interactions in these email exchanges and forum discussions, we propose to apply speech act recognition to email and forum threads.

Despite extensive studies of speech act recognition in many areas, developing speech act recognition for online forms of conversation is very challenging. A major challenge is that emails and forums usually have no labeled data for training statistical speech act recognizers. Fortunately, labeled speech act data are available in other domains (i.e., telephone and meeting conversations in this paper) and large unlabeled data sets can be collected from the Web. Thus, we focus on the problem of how to accurately recognize speech acts in emails and forums by making maximum use of data from existing resources.

Recently, there are increasing interests in speech act recognition of online text-based conversations. Analysis of speech acts for online chat and instant messages and have been studied in computer-mediated communication (CMC) and distance learning (Twitchell et al., 2004; Nastri et al., 2006; Rosé et al., 2008). In natural language processing, Cohen et al. (2004) and Feng et al. (2006) used speech acts to capture the intentional focus of emails and discussion boards. However, they assume that enough labeled data are available for developing speech act recognition models.

A main contribution of this paper is that we address the problem of learning speech act recognition in a semi-supervised way. To our knowledge, this is the first use of semi-supervised speech act recognition in emails and online forums. To do this, we make use of labeled data from spoken conversations (Jurafsky et al., 1997; Dhillon et al., 2004). A second contribution is that our model learns subtree features that constitute discriminative patterns: for example, variable length n-grams and partial dependency structures. Therefore, our model can capture both local features such as n-grams and non-local dependencies. In this paper, we extend subtree pattern mining to the semi-supervised learning problem.

This paper is structured as follows. Section 2 reviews prior work on speech act recognition and Section 3 presents the problem statement and our data sets. Section 4 describes a supervised method of learning subtree features that shows the effectiveness of subtree features on labeled data sets. Section 5 proposes semi-supervised learning techniques for speech act recognition and Section 6 demonstrates our method applied to email and on-

[*]This work was conducted during the author's internship at Microsoft Research Asia.

[1]http://tripadvisor.com/

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1250–1259,
Singapore, 6-7 August 2009.

line forum thread data. Section 7 concludes this paper with future work.

2 Related Work

Speech act theory is fundamental to many studies in discourse analysis and pragmatics (Austin, 1962; Searle, 1969). A speech act is an illocutionary act of conversation and reflects shallow discourse structures of language. Recent research on spoken dialog processing has investigated computational speech act models of human-human and human-computer conversations (Stolcke et al., 2000) and applications of these models to CMC and distance learning (Twitchell et al., 2004; Nastri et al., 2006; Rosé et al., 2008).

Our work in this paper is closely related to prior work on email and forum speech act recognition. Cohen et al. (2004) proposed the notion of 'email speech act' for classifying the intent of an email sender. They defined verb and noun categories for email speech acts and used supervised learning to recognize them. Feng et al. (2006) presented a method of detecting conversation focus based on the speech acts of messages in discussion boards. Extending Feng et al. (2006)'s work, Ravi and Kim (2007) applied speech act classification to detect unanswered questions. However, none of these studies have focused on the semi-supervised speech act recognition problem and examined their methods across different genres.

The speech processing community frequently employs two large-scale corpora for speech act annotation: Switchboard-DAMSL (SWBD) and Meeting Recorder Dialog Act (MRDA). SWBD is an annotation scheme and collection of labeled dialog act[2] data for telephone conversations (Jurafsky et al., 1997). The main purpose of SWBD is to acquire stochastic discourse grammars for training better language models for automatic speech recognition. More recently, an MRDA corpus has been adapted from SWBD but its tag set for labeling meetings has been modified to better reflect the types of interaction in multi-party face-to-face meetings (Dhillon et al., 2004). These two corpora have been extensively studied, e.g., (Stolcke et al., 2000; Ang et al., 2005; Galley et al., 2004). We also use these for our experiments.

[2]A dialog act is the meaning of an utterance at the level of illocutionary force (Austin, 1962), and broadly covers the speech act and adjacency pair (Stolcke et al., 2000). In this paper, we use only the term 'speech act' for clarity.

This paper focuses on the problem of semi-supervised speech act recognition. The goal of semi-supervised learning techniques is to use auxiliary data to improve a model's capability to recognize speech acts. The approach in Tur et al. (2005) presented semi-supervised learning to employ auxiliary unlabeled data in call classification, and is closely related to our work. However, our approach uses the most discriminative subtree features, which is particularly attractive for reducing the model's size. Our problem setting is closely related to the domain adaptation problem (Ando and Zhang, 2005), i.e., we seek to obtain a model that analyzes target domains (emails and forums) by adapting a method that analyzes source domains (SWBD and MRDA). Recently, this type of domain adaptation has become an important topic in natural language processing.

3 Problem Definition

3.1 Problem Statement

We define speech act recognition to be the task that, given a sentence, maps it to one of the speech act types. Figure 1 shows two examples of our email and forum speech act recognition. E1∼6 are all sentences in an email message. F1∼3, F4∼5, and F6 are three posts in a forum thread. A sentence interacts alone or with others, for example, F6 agrees with the previous post (F4∼5). To gain insight into our work, it is useful to consider that E2, 3 and F1, 4, 6 are summaries of two discourses. In particular, F1 denotes a question and F4 and F6 are corresponding answers. More recently, using speech acts has become an appealing approach in summarizing the discussions (Galley et al., 2004; McKeown et al., 2007).

Next, we define speech act category based on MRDA. Dhillon et al. (2004) included definitions of speech acts for colloquial style interactions (e.g., backchannel, disruption, and floorgrabber), but these are not applicable in emails and forums. After removing these categories, we define 12 tags (Table 1). Dhillon et al. (2004) provides detailed descriptions of each tag. We note that our tag set definition is different from (Cohen et al., 2004; Feng et al., 2006; Ravi and Kim, 2007) for two reasons. First, prior work primarily interested in the domain-specific speech acts, but our work use domain-independent speech act tags. Second, we focus on speech act recognition on the sentence-level.

E1: I am planning my schedule at CHI 2003 (http://www.chi2003.org/)	S
E2: - will there be anything happening at the conference related to this W3C User interest group?	QY
E3: I do not see anything on the program yet, but I suspect we could at least have an informal SIG	S
E4: - a chance to meet others and bring someone like me up to speed on what is happening.	S
E5: There will be many competing activities, so the sooner we can set this up the more likely I can attend.	S
E6: Keith	S
F1: If given a choice, should I choose Huangpu area, or should I choose Pudong area?	QR
F2: Both location are separated by a Huangpu river, not sure which area is more convenient for sight seeing?	QW
F3: Thanks in advance for reply!	P
F4: Stay on the Puxi side of the Huangpu river and visit the Pudong side by the incredible tourist tunnel.	AC
F5: If you stay on the Pudong side add half an hour to visit the majority of the tourist attractions.	S
F6: I definitely agree with previous post.	AA

Figure 1: Examples of speech act recognition in emails and online forums. Tags are defined in Table 1.

Table 1: Tags used to describe components of speech acts

Tag	Description
A	Accept response
AA	Acknowledge and appreciate
AC	Action motivator
P	Polite mechanism
QH	Rhetorical question
QO	Open-ended question
QR	Or/or-clause question
QW	Wh-question
QY	Yes-no question
R	Reject response
S	Statement
U	Uncertain response

The goal of semi-supervised speech act recognition is to learn a classifier using both labeled and unlabeled data. We formally define our problem as follows. Let $\mathbf{x} = \{x_j\}$ be a *forest*, i.e., a set of trees that represents a natural language structure, for example, a sequence of words and a dependency parse tree. We will describe this in more detail in Section 4. Let y be a *speech act*. Then, we define $\mathcal{D}_L = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$ as the set of labeled training data, and $\mathcal{D}_U = \{\mathbf{x}_i\}_{i=n+1}^{l}$ as the set of unlabeled training data where $l = n + m$ and m is the number of unlabeled data instances. Our goal is to find a learning method to minimize the classification errors in $\mathcal{D}_L$ and $\mathcal{D}_U$.

3.2 Data Preparation

In this paper, we separate labeled ($\mathcal{D}_L$) and unlabeled data ($\mathcal{D}_U$). First we use SWBD[3] and MRDA[4] as our labeled data. We automatically map original annotations in SWBD and MRDA to one of the 12 speech acts.[5] Inter-annotator agreement κ in both data sets is ~ 0.8 (Jurafsky et al., 1997; Dhillon et al., 2004). For evaluation purposes, we divide labeled data into three sets: training, development, and evaluation sets (Table 2). Of the 1,155 available conversations in the SWBD corpus, we use 855 for training, 100 for development, and 200 for evaluation. Among the 75 available meetings in the MRDA corpus, we exclude two meetings of different natures (btr001 and btr002). Of the remaining meetings, we use 59 for training, 6 for development, and 8 for evaluation. Then we merge multi-segments utterances that belong to the same speaker and then divide all data sets into sentences.

As stated earlier, our unlabeled data consists of email (EMAIL) and online forum (FORUM) data. For the EMAIL set, we selected 22,391 emails from Enron data[6] (discussion_threads, all_documents, and calendar folders). For the FORUM set, we crawled 11,602 threads and 55,743 posts from the TripAdvisor travel forum site (Beijing, Shanghai, and Hongkong forums). As our evaluation sets, we used 40 email threads of the BC3 corpus[7] for EMAIL and 100 threads selected from the same travel forum site for FORUM. Every sentences was automatically segmented by the MSRA sentence boundary detector (Table 2). Annotation was performed by two human annotators, and inter-annotator agreements were $\kappa = 0.79$ for EMAIL and $\kappa = 0.73$ for FORUM.

Overall performance of automatic evaluation measures usually depends on the distribution of tags. In both labeled and unlabeled sets, the most

[3]LDC Catalog No. LDC97S62

[4]http://www.icsi.berkeley.edu/~ees/dadb/

[5]Our mapping tables are available at http://home.postech.ac.kr/~stardust/acl09/.

[6]http://www.cs.cmu.edu/~enron/

[7]http://www.cs.ubc.ca/nest/lci/bc3.html

Table 2: Number of sentences in labeled and unlabeled data

Set	SWBD	MRDA
Training	96,553	50,865
Development	12,299	8,366
Evaluation	24,264	10,492

Set	EMAIL	FORUM
Unlabeled	122,125	297,017
Evaluation	2,267	3,711

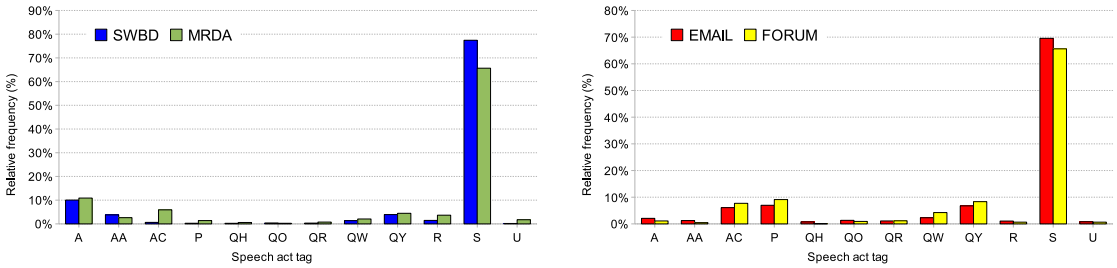


Figure 2: Distribution of speech acts in the evaluation sets. Tags are defined in Table 1.

frequent tag is the statement (S) tag (Figure 2). Distributions of tags are similar in training and development sets of SWBD and MRDA.

4 Speech Act Recognition

Previous work in speech act recognition used a large set of lexical features, e.g., bag-of-words, bigrams and trigrams (Stolcke et al., 2000; Cohen et al., 2004; Ang et al., 2005; Ravi and Kim, 2007). However, these methods create a large number of lexical features that might not be necessary for speech act identification. For example, a Wh-question "What site should we use to book a Beijing-Chonqing flight?" can be predicted by two discriminative features, "(<s>, WRB) → QW" and "(?, </s>) → QW" where <s> and </s> are sentence start and end symbols, and WRB is a part-of-speech tag that denotes a Wh-adverb. In addition, useful features could be of various lengths, i.e. not fixed length n-grams, and non-adjacent. One key idea of this paper is a novel use of subtree features to model these for speech act recognition.

4.1 Exploiting Subtree Features

To exploit subtree features in our model, we use a *subtree pattern mining* method proposed by Kudo and Matsumoto (2004). We briefly introduce this algorithm here. In Section 3.1, we defined $\mathbf{x} = \{x_j\}$ as the forest that is a set of trees. More precisely, x_j is a labeled ordered tree where each node has its own label and is ordered left-to-right. Several types of labeled ordered trees

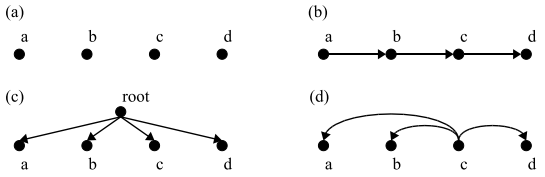


Figure 3: Representations of tree: (a) bag-of-words, (b) n-gram, (c) word pair, and (d) dependency tree. A node denotes a word and a directed edge indicates a parent-and-child relationship.

are possible (Figure 3). Note that S-expression can be used instead for computation, for example `(a(b(c(d))))` for the n-gram (Figure 3(b)). Moreover, we employ a combination of multiple trees as the input of the subtree pattern mining algorithm.

We extract subtree features from the forest set $\{\mathbf{x}_i\}$. A subtree t is a tree if $t \subseteq \mathbf{x}$. For example, `(a)`, `(a(b))`, and `(b(c(d)))` are subtrees of Figure 3(b). We define the subtree feature as a weak learner:

$$f(y, t, \mathbf{x}) \triangleq \begin{cases} +y & t \subseteq \mathbf{x}, \\ -y & \text{otherwise}, \end{cases} \quad (1)$$

where we assume a binary case $y \in \mathcal{Y} = \{+1, -1\}$ for simplicity. Even though the approach in Kudo and Matsumoto (2004) and ours are simiar, there are two clear distinctions. First, our method employs multiple tree structures, and uses different constraints to generate subtree candidates. In this paper, we only restrict generating

the dependency subtrees which should have 3 or more nodes. Second, our method is of interest for semi-supervised learning problems. To learn subtree features, Kudo and Matsumoto (2004) assumed supervised data $\{(\mathbf{x}_i, y_i)\}$. Here, we describe the supervised learning method and will describe our semi-supervised method in Section 5.

4.2 Supervised Boosting Learning

Given training examples, we construct a ensemble learner $F(\mathbf{x}) = \sum_k \lambda_k f(y_k, t_k, \mathbf{x})$, where λ_k is a coefficient for linear combination. A final classifier $h(\mathbf{x})$ can be derived from the ensemble learner, i.e., $h(\mathbf{x}) \triangleq sgn(F(\mathbf{x}))$. As an optimization framework (Mason et al., 2000), the objective of boosting learning is to find F such that the cost of functional

$$\mathcal{C}(F) = \sum_{i \in \mathcal{D}} \alpha_i C[y_i F(\mathbf{x}_i)] \tag{2}$$

is minimized for some non-negative and monotonically decreasing cost function $C : \mathbb{R} \to \mathbb{R}$ and the weight $\alpha_i \in \mathbb{R}^+$. In this paper, we use the AdaBoost algorithm (Schapire and Singer, 1999); thus the cost function is defined as $C(z) = e^{-z}$.

Constructing an ensemble learner requires that the user choose a base learner, $f(y, t, \mathbf{x})$, to maximize the inner product $-\langle \nabla \mathcal{C}(F), f \rangle$ (Mason et al., 2000). Finding $f(y, t, \mathbf{x})$ to maximize $-\langle \nabla \mathcal{C}(F), f \rangle$ is equivalent to searching for $f(y, t, \mathbf{x})$ to minimize $2 \sum_{i: f(y,t,\mathbf{x}_i) \neq y_i} w_i - 1$, where w_i for $i \in \mathcal{D}_L$, is the empirical data distribution $w_i^{(k)}$ at step k. It is defined as:

$$w_i^{(k)} = \alpha_i \cdot e^{-y_i F(\mathbf{x}_i)}. \tag{3}$$

From Eq. 3, a proper base learner (i.e., subtree) can be found by maximizing weighted gain, where

$$\text{gain}(t, y) = \sum_{i \in \mathcal{D}_L} y_i w_i f(y, t, \mathbf{x}_i). \tag{4}$$

Thus, subtree mining is formulated as the problem of finding $(\hat{t}, \hat{y}) = \underset{(t,y) \in \mathcal{X} \times \mathcal{Y}}{\arg\max}\ \text{gain}(t, y)$. We need to search with respect to a non-monotonic score function (Eq. 4), thus we use the monotonic bound, $\text{gain}(t, y) \leq \mu(t)$, where

$$\mu(t) = \max \left(2 \sum_{\{i | y_i = +1, t \subseteq \mathbf{x}_i\}} w_i - \sum_{i=1}^{n} y_i f(y, t, \mathbf{x}_i), \right.$$
$$\left. 2 \sum_{\{i | y_i = -1, t \subseteq \mathbf{x}_i\}} w_i + \sum_{i=1}^{n} y_i f(y, t, \mathbf{x}_i) \right). \tag{5}$$

Table 3: Result of supervised learning experiment; columns are micro-averaged F_1 score with macro-averaged F_1 score in parentheses. MAXENT: maximum entropy model; BOW: bag-of-words model; NGRAM: n-gram model; +POSTAG, +DEPTREE, +SPEAKER indicate that the components were added individually onto NGRAM. '*' indicates results significantly better than the NGRAM model ($p < 0.001$).

Model	SWBD	MRDA
MAXENT	92.76 (63.54)	82.48 (57.19)
BOW	91.32 (54.47)	82.17 (55.42)
NGRAM	92.60 (58.43)	83.30 (57.53)
+POSTAG	92.69 (60.07)	83.60 (58.46)
+DEPTREE	92.67 (61.75)	*83.57 (57.45)
+SPEAKER	*92.86 (63.13)	83.40 (58.20)
ALL	*92.87 (63.77)	83.49 (59.04)

The subtree set is efficiently enumerated using a branch-and-bound procedure based on $\mu(t)$ (Kudo and Matsumoto, 2004).

After finding an optimal base leaner, $f(\hat{y}, \hat{t}, \mathbf{x})$, we need to set the coefficient λ_k to form a new ensemble, $F(\mathbf{x}_i) \leftarrow F(\mathbf{x}_i) + \lambda_k f(\hat{t}, \hat{y}, \mathbf{x}_i)$. In AdaBoost, we choose

$$\lambda_k = \frac{1}{2} \log \left(\frac{1 + \text{gain}(\hat{t}, \hat{y})}{1 - \text{gain}(\hat{t}, \hat{y})} \right). \tag{6}$$

After K iterations, the boosting algorithm returns the ensemble learner $F(\mathbf{x})$ which consists of a set of appropriate base learners $f(y, t, \mathbf{x})$.

4.3 Evaluation on Labeled Data

We verified the effectiveness of using subtree features on the SWBD and MRDA data sets. For boosting learning, one typically assumes $\alpha_i = 1$. In addition, the number of iterations, which relates to the number of patterns, was determined by a development set. We also used a one-vs.-all strategy for the multi-class problem. Precision and recall were computed and combined into micro- and macro-averaged F_1 scores. The significance of our results was evaluated using the McNemar paired test (Gillick and Cox, 1989), which is based on individual labeling decisions to compare the correctness of two models. All experiments were implemented in C++ and executed in Windows XP on a PC with a Dual 2.1 GHz Intel Core2 processor and 2.0 Gbyte of main memory.

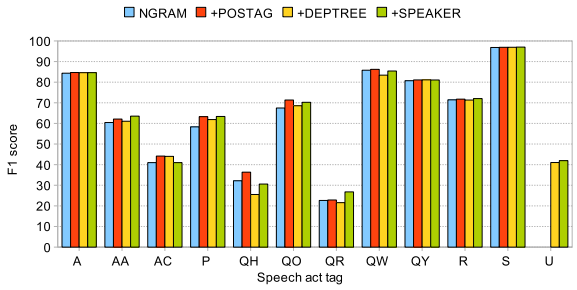


Figure 4: Comparison of different trees (SWBD)

We show that use of subtree features is effective to solve the supervised speech act recognition problem. We also compared our model with the state-of-the-art maximum entropy classifier (MAXENT). We used bag-of-words, bigram and trigram features for MAXENT, which modeled 702k (SWBD) and 460k (MRDA) parameters (i.e., patterns), and produced micro-averaged F_1 scores of 92.76 (macro-averaged F_1 = 63.54) for SWBD and 82.48 (macro-averaged F_1 = 57.19) for MRDA. In contrast, our method generated approximately 4k to 5k patterns on average with similar or greater F_1 scores (Table 3); hence, compared to MAXENT, our model requires fewer calculations and is just as accurate.

The n-gram model (NGRAM) performed significantly better than the bag-of-words model (McNemar test; $p < 0.001$) (Table 3). Unlike MAXENT, NGRAM automatically selects a relevant set of variable length n-gram features (i.e., phrase features). To this set, we separately added two syntax type features, part-of-speech tag n-gram (POSTAG) and dependency parse tree (DEPTREE) automatically parsed by Minipar[8], and one discourse type feature, speaker n-gram (SPEAKER). Although some micro-averaged F_1 are not statistically significant between the original NGRAM and the models that include POSTAG, DEPTREE or SPEAKER, macro-averaged F_1 values indicate that minor classes can take advantage of other structures. For example, in the result of SWBD (Figure 4), DEPTREE and SPEAKER models help to predict uncertain responses (U), whereas NGRAM and POSTAG cannot do this.

5 Semi-supervised Learning

Our goal is to eventually make maximum use of existing resources in SWBD and MRDA for email/forum speech act recognition. We call the model trained on the mixed data of these two corpora BASELINE. We use ALL features in constructing the BASELINE for the semi-supervised experiments. While this model gave promising results using SWBD and MRDA, language used in emails and forums differs from that used in spoken conversation. For example, 'thanx' is an expression commonly used as a polite mechanism in online communications. To adapt our model to understand this type of difference between spoken and online text-based conversations, we should induce new patterns from unlabeled email and forum data. We describe here two methods of semi-supervised learning.

5.1 Method 1: Bootstrapping

First, we bootstrap the BASELINE model using automatically predicted unlabeled examples. However, using all of the unlabeled data results in noisy models; therefore filtering or selecting data is very important in practice. To this end, we only select similar examples by criterion, $d(\mathbf{x}_i, \mathbf{x}_j) < r$ or k nearest neighbors where $\mathbf{x}_i \in \mathcal{D}_L$ and $\mathbf{x}_j \in \mathcal{D}_U$. In practice, r or k are fixed. In our method, examples are represented by trees; hence we use a "tree edit distance" for calculating $d(\mathbf{x}_i, \mathbf{x}_j)$ (Shasha and Zhang, 1990). Selected examples are evaluated using BASELINE, and using subtree pattern mining runs on the augmented data (i.e. unlabeled). We call this method BOOTSTRAP.

5.2 Method 2: Semi-supervised Boosting

Our second method is based on a principle of semi-supervised boosting learning (Bennett et al., 2002). Because we have no supervised guidance for $\mathcal{D}_U$, our objective functional to find F is defined as:

$$\mathcal{C}(F) = \sum_{i \in \mathcal{D}_L} \alpha_i C[y_i F(\mathbf{x}_i)] + \sum_{i \in \mathcal{D}_U} \beta_i C[|F(\mathbf{x}_i)|] \quad (7)$$

This cost functional is non-differentiable. To solve it, we introduce pseudo-labels $\tilde{y}$ where $\tilde{y} = sgn(F(\mathbf{x}))$ and $|F(\mathbf{x})| = \tilde{y}F(\mathbf{x})$. Using the same derivation in Section 4.2, we obtain the following

[8] http://www.cs.ualberta.ca/∼lindek/minipar.htm

gain function and update rules:

$$\text{gain}(t, y) = \sum_{i \in \mathcal{D}_L} y_i w_i f(y, t, \mathbf{x}_i) + \sum_{i \in \mathcal{D}_U} \tilde{y}_i w_i f(y, t, \mathbf{x}_i), \quad (8)$$

$$w_i = \begin{cases} \alpha_i \cdot e^{-y_i F(\mathbf{x}_i)} & i \in \mathcal{D}_L, \\ \beta_i \cdot e^{-\tilde{y}_i F(\mathbf{x}_i)} & i \in \mathcal{D}_U. \end{cases} \quad (9)$$

Intuitively, an unlabeled example that has a high-confidence $|F(\mathbf{x})|$ at the current step, will probably receive more weight at the next step. That is, similar instances become more important when learning and mining subtrees. This semi-supervised boosting learning iteratively generates pseudo-labels for unlabeled data and finds the value of F that minimizes training errors (Bennett et al., 2002). Also, the algorithm infers new features from unlabeled data, and these features are iteratively re-evaluated by the current ensemble learner. We call this method SEMIBOOST.

6 Experiment

6.1 Setting

We describe specific settings used in our experiment. Because we have no development set, we set the maximum number of iterations K at 10,000. At most K patterns can be extracted, but this seldom happens because duplicated patterns are merged. Typical settings for semi-supervised boosting are $\alpha_i = 1$ and $\beta_i = 0.5$, that is, we penalize the weights for unlabeled data.

For efficiency, BASELINE model used 10% of the SWBD and MRDA data, selected at random. We observed that this data set does not degrade the results of semi-supervised speech act recognition. For BOOTSTRAP and SEMIBOOST, we selected $k = 100$ nearest neighbors of unlabeled examples for each labeled example using tree edit distance, and then used 24,625 (SWBD) and 54,961 (MRDA) sentences for the semi-supervised setting.

All trees were combined as described in Section 4.3 (ALL model). In EMAIL and FORUM data we added different types of discourse features: message type (e.g., initial or reply posts), authorship (e.g., an identification of 2nd or 3rd posts written by the same author), and relative position of a sentence. In Figure 1, for example, F1∼3 is an initial post, and F4∼5 and F6 are reply posts. Moreover, F1, F4, and F6 are the first sentence in each post.

Table 4: Results of speech act recognition on online conversations; columns are micro-averaged F_1 score with macro-averaged scores in parentheses. '⋆' indicates that the result is significantly better than BASELINE ($p < 0.001$).

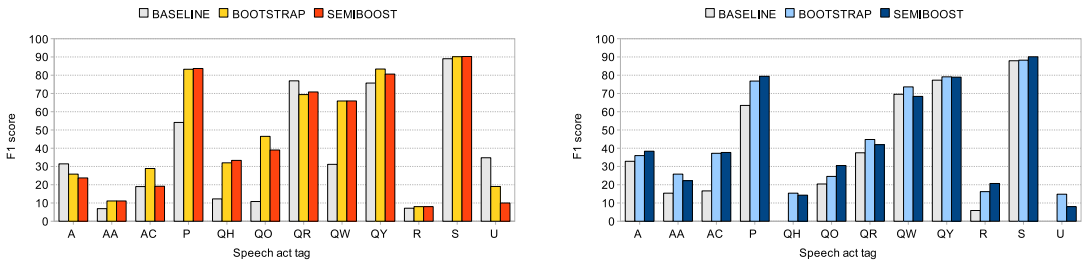
Model	EMAIL	FORUM
BASELINE	78.87 (37.44)	78.93 (35.57)
BOOTSTRAP	⋆83.11 (44.90)	79.09 (44.38)
SEMIBOOST	⋆82.80 (44.64)	⋆81.76 (44.21)
SUPERVISED	90.95 (75.71)	83.67 (40.68)

These features do not occur in SWBD or MRDA because these are utterance-by-utterance conversations.

6.2 Result and Discussion

First, we show that our method of semi-supervised learning can improve modeling of the speech act of emails and forums. As our baseline, BASELINE achieved a micro-averaged F_1 score of $\sim$ 79 for both data sets. This implies that SWBD and MRDA data are useful for our problem. Using unlabeled data, semi-supervised methods BOOTSTRAP and SEMIBOOST perform better than BASELINE (Table 4; Figure 5). To verify our claim, we evaluated the supervised speech act recognition on EMAIL and FORUM evaluation sets with 5-fold cross validation (SUPERVISED in Table 4). In particular, our semi-supervised speech act recognition is competitive with the supervised model in FORUM data.

The difference in performance between supervised results in EMAIL and FORUM seems to indicate that the latter is a more difficult data set. However, our SEMIBOOST method were able to come close to the supervised FORUM results (81.76 vs. 83.67). This is also close to the range of supervised MRDA data set ($F_1 = 83.49$ for ALL, Table 3). Moreover, we analyzed a main reason of why transfer results were competitive in the FORUM but not in the EMAIL. This might be due to the mismatch in the unlabeled data, that is, we used different email collections, the BC3 corpus (email communication of W3C on w3.org sites), for evaluation while used Enron data for adaption. We also conjecture that the discrepancy between EMAIL and FORUM is probably due to the more heterogeneous nature of the FORUM data where anyone can post and reply while EMAIL (Enron or



(a) EMAIL (b) FORUM

Figure 5: Result of the semi-supervised learning method

BC3) might have a more fix set of participants.

The improvement of less frequent tags is prominent, for example 25% for action motivator (AC), 40% for polite mechanism (P), and 15% for rhetorical question (QR) error rate reductions were achieved in FORUM data (Figure 5(b)). Therefore, the semi-supervised learning method is more effective with small amounts of labeled data (i.e., less frequent annotations). We believe that despite their relative rarity, these speech acts are more important than the statement (S) in some applications, e.g., summarization.

Next, we give a qualitative analysis for better interpretation of our problem and results. Due to limited space, we focus on FORUM data, which can potentially be applied to many applications. Of the top ranked patterns extracted by SEMIBOOST (Figure 6(a)), subtree patterns of n-gram, part-of-speech, dependency parse trees are most discriminative. The patterns from unlabeled data have relatively lower ranks, but this is not surprising. This indicates that BASELINE model provides the base knowledge for semi-supervised speech act recognition. Also, unlabeled data for EMAIL and FORUM help to induce new patterns or adjust the model's parameters. As a result, the semi-supervised method is better than the BASELINE when an identical number of patterns is modeled (Figure 6(b)). For this result, we conclude that our method successfully transfers knowledge from a source domain (i.e., SWBD and MRDA) to a target domain (i.e., EMAIL and FORUM); hence it can be a solution to the domain adaption problem.

Finally, we determine the main reasons for error (in SEMIBOOST), to gain insights that may allow development of better models in future work (Figure 6(c)). We sorted speech act tags by their semantics and partitioned the confusion matrix into question type (Q*) and statement, which are two high-level speech acts. Most errors occur in the similar categories, that is, language usage in question discourse is definitely distinct from that in statement discourse. From this analysis, we believe that more advanced techniques (e.g. two-stage classification and learning with hierarchy-augmented loss) can improve our model.

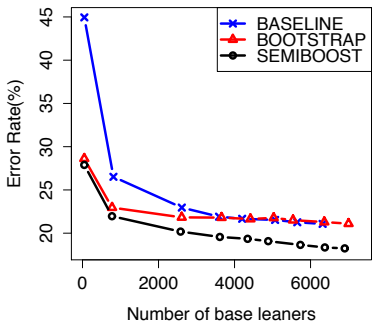
7 Conclusion

Despite the increasing interest in online text-based conversations, no study to date has investigated semi-supervised speech act recognition in email and forum threads. This paper has addressed the problem of learning to recognize speech acts using labeled and unlabeled data. We have also contributed to the development of a novel application of boosting subtree mining. Empirical results have demonstrated that semi-supervised learning of speech act recognition with subtree features improves the performance in email and forum data sets. An attractive future direction is to exploit prior knowledge for semi-supervised speech act recognition. Druck et al. (2008) described generalized expectation criteria in which a discriminative model can employ the labeled features and unlabeled instances. Using prior knowledge, we expect that our model will effectively learn useful patterns from unlabeled data.

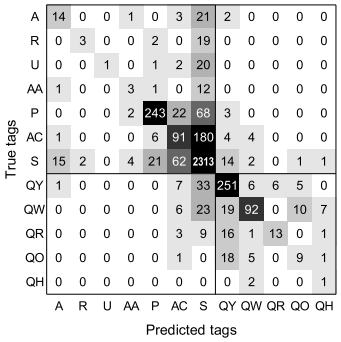
As work progresses on analyzing online text-based conversations such as emails, forums, and online chats, the importance of developing models for discourse without annotating much new data will become more important. In the future, we plan to explore other related problems such as adjacency pairs (Levinson, 1983) and discourse parsing (Soricut and Marcu, 2003) for large-scale online forum data.

Tag	Example pattern
A	(ROOT (yep))
AA	(<s> (wow (. (</s>))))
AC	(WRB (VB (NN (PRP))))
P	(thanks)
QH	(cares (?))
QO	(ROOT (think) (?))
QR	(ROOT (or) (?))
QW	(ROOT (rel=sub (what)))
QY	(<s> (do))
R	(nay)
S	(ROOT (U (,)) (?))
U	(it (is (possible (.))))

(a) Example patterns



(b) Learning behavior



(c) Confusion matrix

Figure 6: Analysis on FORUM data

Acknowledgement

We would like to thank to anonymous reviewers for their valuable comments, and Yunbo Cao, Wei Lai, Xinying Song, Jingtian Jing, and Wei Wu for their help in preparing our data.

References

R. Ando and T. Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

J. Ang, Y. Liu, and E. Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of ICASSP*, pages 1061–106.

J. Austin. 1962. *How to Do Things With Words*. Harvard Univ. Press, Cambridge, MA.

K.P. Bennett, A. Demiriz, and R. Maclin. 2002. Exploiting unlabeled data in ensemble methods. In *Proceedings of ACM SIGKDD*, pages 289–296.

W.W. Cohen, V.R. Carvalho, and T. Mitchell. 2004. Learning to classify email into "speech acts". In *Proceedings of EMNLP*, pages 309–316.

R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg. 2004. Meeting recorder project: Dialog act labeling guide. Technical report, International Computer Science Institute.

G. Druck, G. Mann, and A. McCallum. 2008. Learning from labeled features using generalized expectation criteria. In *Proceedings of ACM SIGIR*, pages 595–602.

D. Feng, E. Shaw, J. Kim, and E. H. Hovy. 2006. Learning to detect conversation focus of threaded discussions. In *Proceedings of HLT-NAACL*, pages 208–215.

M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg. 2004. Identifying agreement and disagreement in conversational speech: use of bayesian networks to model pragmatic dependencies. In *Proceedings of ACL*.

L. Gillick and S. Cox. 1989. Some statistical issues in the comparison of speech recognition algorithms. In *Proceedings of ICASSP*, pages 532–535.

D. Jurafsky, E. Shriberg, and D. Biasca. 1997. Switchboard SWBD-DAMSL labeling project coder's manual, draft 13. Technical report, Univ. of Colorado Institute of Cognitive Science.

T. Kudo and Y. Matsumoto. 2004. A boosting algorithm for classification of semi-structured text. In *Proceedings of EMNLP*, pages 301–308.

S. Levinson. 1983. *Pragmatics*. Cambridge Univ. Press, Cambridge.

L. Mason, P. Bartlett, J. Baxter, and M. Frean. 2000. Functional gradient techniques for combining hypotheses. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 221–246. MIT Press, Cambridge, MA.

K. McKeown, L. Shrestha, and O. Rambow. 2007. Using question-answer pairs in extractive summarization of email conversations. In *Proceedings of CICLing*, volume 4394 of *Lecture Notes in Computer Science*, pages 542–550.

J. Nastri, J. Pe na, and J. T. Hancock. 2006. The construction of away messages: A speech act analysis. *Journal of Computer-Mediated Communication*, 11(4):article 7.

S. Ravi and J. Kim. 2007. Profiling student interactions in threaded discussions with speech act classifiers. In *Proceedings of the AI in Education Conference*.

C. Rosé, Y. Wang, Y. Cui, J. Arguello, K. Stegmann, A. Weinberger, and F. Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, 3(3):237–271.

R.E. Schapire and Y. Singer. 1999. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336.

J. Searle. 1969. *Speech Acts*. Cambridge Univ. Press, Cambridge.

D. Shasha and K. Zhang. 1990. Fast algorithms for the unit cost editing distance between trees. *Journal of Algorithms*, 11(4):581–621.

R. Soricut and D. Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of NAACL-HLT*, pages 149–156.

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

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

D. P. Twitchell, J. F. Nunamaker, and J. K. Burgoon. 2004. Using speech act profiling for deception detection. In *Second Symposium on Intelligence and Security Informatics*, volume 3073 of *Lecture Notes in Computer Science*, pages 403–410.