Towards Scalable Speech Act Recognition in Twitter: Tackling Insufficient Training Data

Renxian Zhang **Dehong Gao** **Wenjie Li**
Department of Computing
The Hong Kong Polytechnic University
{csrzhang, csdgao, cswjli}@comp.polyu.edu.hk

Abstract

Recognizing speech act types in Twitter is of much theoretical interest and practical use. Our previous research did not adequately address the deficiency of training data for this multi-class learning task. In this work, we set out by assuming only a small seed training set and experiment with two semi-supervised learning schemes, transductive SVM and graph-based label propagation, which can leverage the knowledge about unlabeled data. The efficacy of semi-supervised learning is established by our extensive experiments, which also show that transductive SVM is more suitable than graph-based label propagation for our task. The empirical findings and detailed evidences can contribute to scalable speech act recognition in Twitter.

1. Introduction

The social media platform of Twitter makes available a plethora of data to probe the communicative act of people in a social network woven by interesting events, people, topics, etc. Communicative acts such as disseminating information, asking questions, or expressing feelings all fall in the purview of "speech act", a long established area in pragmatics (Austin 1962). The automatic recognition of speech act in tons of tweets has both theoretical and practical appeal. Practically, it helps tweeters to find topics to read or tweet about based on speech act compositions. Theoretically, it introduces a new dimension to study social media content as well as providing real-life data to validate or falsify claims in the speech act theory.

Different taxonomies of speech act have been proposed by linguists and computational linguists, ranging from a few to over a hundred types. In this work, we adopt the 5 types of speech act used in our previous work (Zhang et al. 2011), which are in turn inherited from (Searle 1975): **statement**, **question**, **suggestion**, **comment**, and **miscellaneous**. Our choice is based on the fact that unlike face-to-face communication, twittering is more in a "broadcasting" style than on a personal basis. Statement and comment, which are usually intended to make one's knowledge, thought, and sentiment known, thus befit Twitter's communicative style. Question and suggestion on Twitter are usually targeted at other tweeters in general or one's followers. More interpersonal speech acts such as "threat" or "thank" as well as rare speech acts in Twitter (Searle's (1975) "commissives" and "declaratives") are relegated to "miscellaneous". Some examples from our experimental datasets are provided in Table 1.

Tweet	*Speech Act*
Libya Releases 4 Times Journalists - http://www.photozz.com/?104k	**Statement**
#sincewebeinghonest why u so obsessed with what me n her do?? Don't u got ya own man???? Oh wait.....	**Question**
RT @NaonkaMixon: I will donate 10 $ to the Red Cross Japan Earthquake fund for every person that retweets this! #PRAYFORJAPAN	**Suggestion**
is enjoying this new season of #CelebrityApprentice.... Nikki Taylor = Yum!!	**Comment**
65. I want to get married to someone i meet in highschool. #100factsaboutme	**Miscellaneous**

Table 1. Example Tweets with Speech acts

Assuming one tweet demonstrates only one speech act, the automatic recognition of those speech act types in Twitter is a multi-class classification task. We concede that this assumption may not always hold in real situations. But given the short length of tweets, multi-speech act tweets are rare and we find this simplifying assumption effective in reducing the complexity of our problem. A major problem with this task is the deficiency of training data. Tweeters as well as face-to-face interlocutors do not often identify their speech acts; human annotation is costly and time-consuming. Although our previous research (Zhang et al. 2011) sheds light on the preparation of training data, it did not adequately address this problem.

Our contribution in this work is to directly address the problem of training data deficiency by using two well-known semi-supervised learning techniques that leverage the relationship between a small seed of training data and a large body of unlabeled data: transductive SVM and graph-based label propagation. The empirical results show that the knowledge about unlabeled data provides promising solutions to the data deficiency problem, and that transductive SVM is more competent for our task. Our exploration with different training/unlabeled data ratios for three major Twitter categories and a mixed-type category provides solid evidential support for future research.

The rest of the paper is organized as follows. Section 2 reviews works related to speech act recognition and semi-supervised learning; Section 3 briefly discusses supervised learning of speech act types developed in our earlier work and complementing the previous findings with learning curves. The technical details of semi-supervised learning are presented in Section 4. Then we report and discuss the results of our experiments in Section 5. Finally, Section 6 concludes the paper and outlines future directions.

2. Related Work

The automatic recognition of speech act, also known as “dialogue act”, has attracted sustained interest in computational linguistics and speech technology for over a decade (Searle 1975; Stolcke et al. 2000). A few annotated corpora such as Switchboard-DAMSL (Jurafsky et al. 1997) and Meeting Recorder Dialog Act (Dhillon et al. 2004) are widely used, with data transcribed from telephone or face-to-face conversation.

Prior to the flourish of microblogging services such as Twitter, speech act recognition has been extended to electronic media such as email and discussion forum (Cohen et al. 2004; Feng et al. 2006) in order to study the behavior of email or message senders.

The annotated corpora for ordinary verbal communications and the methods developed for email, or discussion forum cannot be directly used for our task because Twitter text has a distinctive Netspeak style that is situated between speech and text but resembles neither (Crystal 2006, 2011). Compared with email or forum post, it is rife with linguistic noises such as spelling mistakes, random coinages, mixed use of letters and symbols.

Speech act recognition in Twitter is a fairly new task. In our pioneering work (Zhang et al. 2011), we show that Twitter text normalization is unnecessary and even counterproductive for this task. More importantly, we propose a set of useful features and draw empirical conclusion about the scope of this task, such as recognizing speech act on the coarse-grade category level works as well as on the fine-grade topic level. In this work, we continue to adopt this framework including other learning details (speech act types and feature selection for tweets), but the new quest starts where the old one left: tackling insufficient training data.

As in many practical applications, sufficient annotated data are hard to obtain. Therefore, unsupervised and semi-supervised learning methods are actively pursued. While unsupervised sentence classification is rule-based and domain-dependent (Deshpande et al. 2010), semi-supervised methods that both alleviate the data deficiency problem and leverage the power of state-of-the-art classifiers hold more promises for different domains (Medlock and Briscoe 2007; Erkan et al. 2007).

In the machine learning literature, a classic semi-supervised learning scheme is proposed by Yarowsky (1995), which is a classical self-teaching process that makes no use of labeled data before they are classified. More theoretical analyses are made by (Culp and Michailidis 2007) and (Haffari and Sarkar 2007).

Transductive SVM (Joachims 1999) extends the state-of-the-art inductive SVM by explicitly considering the relationship between labeled and unlabeled data. The graph-based label propagation model (Zhu et al. 2003; Zhou et al. 2004) using a harmonic function also accommodates the knowledge about unlabeled data. We will adapt both of them to our multi-class classification task.

Jeong et al. (2009) report a semi-supervised approach to classifying speech acts in emails and online forums. But their subtree-based method is not applicable to our task because Twitter's noisy textual quality cannot be found in the much cleaner email or forum texts.

3. Supervised Learning of Speech Act Types

Supervised learning of speech act types in Twitter relies heavily on a good set of features that capture the textual characteristics of both Twitter and speech act utterances. As in our previous work, we use speech act-specific cues, special words (abbreviations and acronyms, opinion words, vulgar words, and emoticons), and special characters (Twitter-specific characters and a few punctuations). Tweet-external features such as tweeter profile may also help, but that is beyond the focus of this paper.

Although it has been empirically shown that speech act recognition in Twitter can be done without using training data specific to topics or even categories, it is not clear how much training data is needed to achieve desirable performance. In order to answer this question, we adopt the same experimental setup and datasets as reported in (Zhang et al. 2011) and plot the learning curves shown in Figure 1.

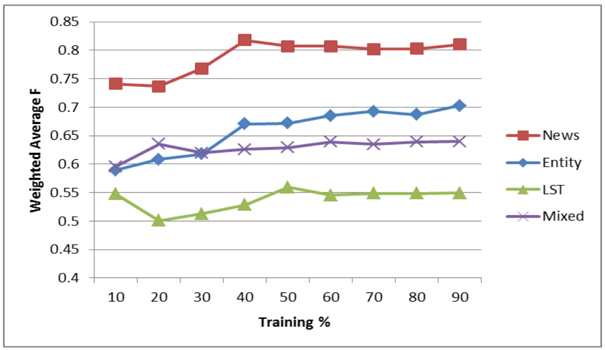

Figure 1. Learning Curves of Each Category and All Tweets

For all individual experiments, the test data are a randomly sampled 10% set of all annotated data. When training data reach 90%, we actually duplicate the reported results. However, Figure 1 shows that it is unnecessary to use so much training data to achieve good classification performance. For News and Entity, the classification makes little noticeable improvement after the training data ratio reaches 40% (training : test = 4 : 1). For Mixed (the aggregate of the News, Entity, LST datasets) and LST, performance peaks even earlier at 20% training data (training : test = 2 : 1) and 10% (training : test = 1 : 1).

It is delightful to see that only a moderate number of annotated data are needed for speech act recognition. But even that number (for the Mixed dataset, 10% training data are over 800 annotated tweets) may not be available and in many situations, test data may be much more than training data. Taking this challenge is the next important step we make.

4. Semi-Supervised Learning of Speech Act Types

The problem setting of a small seed training (labeled) set and a much larger test (labeled) set fits the semi-supervised learning scheme. Classic semi-supervised learning approaches such as self-teaching methods (e.g., Yarowsky 1995) are mainly concerned with incrementing high-confidence labeled data in each round of training. They do not, however, directly take into account the knowledge about unlabeled data. The recent research emphasis is on leveraging knowledge about unlabeled data during training. In this section, we discuss two such approaches.

4.1 Transductive SVM

The standard SVM classifier popularly used in text classification is also known as inductive SVM as a model is induced from training data. The model is solely dependent on the training data and agnostic about the test data. In contrast, transductive SVM (Vapnik 1998; Joachims 1999) predicts test labels by using the knowledge about test data. In the case of test (unlabeled) data far outnumbering training (labeled) data, transductive SVM provides a feasible scheme of semi-supervised learning.

For a single-class classification problem $\{\mathbf{x}_i, y_\mathrm{i}\}$ that focuses on only one speech act type, where $\mathbf{x}_i$ is the *i*th tweet and y_i is the corresponding label and $y_i \in \{+1, -1\}$ denotes whether $\mathbf{x}_i$ contains the speech act or not, inductive SVM is formulated to find an optimal hyperplane $sign(\mathbf{w}\cdot\mathbf{x}_i - b)$ to maximize the soft margin between positive and negative objects, or to minimize:

$$1/2\|\mathbf{w}\|^2 + C\sum_i \phi_i$$
$$\text{s.t. } y_i(\mathbf{x}_i \cdot \mathbf{w} - b) \geq 1 - \phi_i,\ \phi_i \geq 0$$

where ϕ_i is a slack variable. Adopting the same formulation, transductive SVM further considers test data $\mathbf{x}_i$* during training by finding a labeling y_j* and a hyperplane to maximize the soft margin between both training and test data, or to minimize:

$$1/2\|\mathbf{w}\|^2 + C_1\sum_i \phi_i + C_2\sum_i \varphi_i$$
$$\text{s.t. } y_i(\mathbf{x}_i \cdot \mathbf{w} - b) \geq 1 - \phi_i,\ \phi_i \geq 0$$
$$y_i^*(\mathbf{x}_i^* \cdot \mathbf{w} - b) \geq 1 - \varphi_i,\ \varphi_i \geq 0$$

where φ_i is a slack variable for the test data. In fact, labeling test data is done during training.

As the maximal margin approach proves very effective for text classification, its transductive variant that effectively uses the knowledge about test data holds promises of handling the deficiency of labeled data.

4.2 Graph-based Label Propagation

An alternative way of using unlabeled data in semi-supervised learning is based on the intuition that similar objects should belong to the same class, which can be translated into label smoothness on a graph with weights indicating object similarities. This is the idea underlying Zhu et al.'s (2003) graph-based label propagation model using Gaussian random fields.

We again focus on a single-class classification problem. Formally, $\{\mathbf{x}_1, \dots \mathbf{x}_N\}$ are N tweets, having their actual speech act labels $\mathbf{y} = \{y_1, \dots y_L, \dots y_N\}$ ($y_i \in \{1, 0\}$ denoting whether $\mathbf{x}_i$ contains the speech act or not) with the first L of them known, and $\mathbf{f} = \{f_1, \dots f_L, \dots f_N\}$ are their predicted labels. Let $L = \{\mathbf{x}_1, \dots \mathbf{x}_L\}$ and $U = \{\mathbf{x}_{L+1}, \dots \mathbf{x}_N\}$ and the task is to determine $\{f_{L+1}, \dots f_N\}$ for U. We further define a graph $G = (V, E)$, where $V = L \cup U$ and E is weighted by $\mathbf{W} = [w_{ij}]_{N\times N}$ with w_{ij} denoting the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. Preferring label smoothness on G and preserving the given labels, we want to minimize the loss function:

$$E(\mathbf{f}) = 1/2\sum_{i,j\in L\cup U} w_{ij}(f_i - f_j)^2 = \mathbf{f}^T\mathbf{\Delta f}$$
$$\text{s.t. } f_i = y_i\ (i = 1, \dots, L)$$

where $\mathbf{\Delta} = \mathbf{D} - \mathbf{W}$ is the combinatorial graph Laplacian with $\mathbf{D}$ being a diagonal matrix $[d_{ij}]_{N\times N}$ and $d_{ii} = \sum_j w_{ij}$.

This can be expressed as a harmonic function, $h = \operatorname{argmin}_{f_L = y_L} E(\mathbf{f})$, which satisfies the smoothness property on the graph: $h(i) = 1/d_{ii}\sum_k w_{ik}(h(k))$. If we define $p_{ij} = w_{ij} / \sum_k w_{ik}$ and collect p_{ij} and $h(i)$ into matrix $\mathbf{P}$ and column vector $\mathbf{h}$, solving $\Delta h = 0$ s.t. $\mathbf{h}_L = \mathbf{y}_L$ is equivalent to solving $\mathbf{h} = \mathbf{Ph}$.

To find the solution, we can use L and U to partition $\mathbf{h}$ and $\mathbf{P}$:

$$\mathbf{h} = \begin{bmatrix} \mathbf{h}_L \\ \mathbf{h}_U \end{bmatrix}, \mathbf{P} = \begin{bmatrix} \mathbf{P}_{LL}, \mathbf{P}_{LU} \\ \mathbf{P}_{UL}, \mathbf{P}_{UU} \end{bmatrix}$$

and it can be shown that $\mathbf{h}_U = (\mathbf{I} - \mathbf{P}_{UU})^{-1}\mathbf{P}_{UL}\mathbf{y}_L$. To get the final classification result, those elements in $\mathbf{h}_U$ that are greater than a threshold (0.5) become 1 and the others become 0.

This approach propagates labels from labeled data to unlabeled data on the principle of label smoothness. If the assumption about similar tweets having same speech acts holds, it should work well for our problem.

4.3 Multi-class Classification

In the previous formulations, we emphasized "single-class classification" because both transductive SVM and graph-based label propagation are inherently one class-oriented. Since our problem is a multi-class one, we transform the problem to single-class classifications by using the one-vs-all scheme.

Specifically, for each class (speech act type) c_i, we label all training instances belonging to c_i as +1 and all those belonging to other classes as −1 and then do binary classification. For our problem with 5 speech act types, we make 5 such transformations. The final prediction is made by choosing the class with the highest classification score from the 5 binary classifiers. Both transductive SVM and graph-based label propagation produce real-valued classification scores and are amenable to this scheme.

5. Experiments

Our experiments are designed to answer two questions: 1) How useful is semi-supervised speech act learning in comparison with supervised learning? 2) Which semi-supervised learning approach is more appropriate for our problem?

5.1 Experimental Setup

We use the 6 datasets in our previous study[1], which fall into 3 categories: *News*, *Entity*, *Long-standing Topic* (*LST*). Each of the total 8613 tweets is labeled with one of the following speech act types: *sta* (statement), *que* (question), *sug* (suggestion), *com* (comment), *mis* (miscellaneous). In addition, we randomly select 1000 tweets from each of the categories to create a *Mixed* category of 3000 tweets. Figures 2 to 5 illustrate the distributions of the speech act types in the 3 original categories and the *Mixed* category.

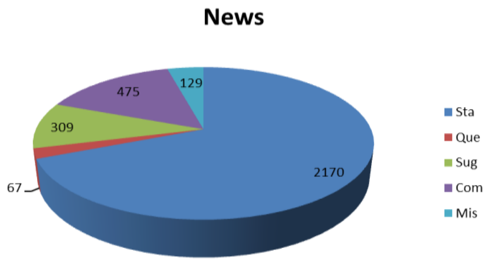

Figure 2. Speech Act Distribution (News)

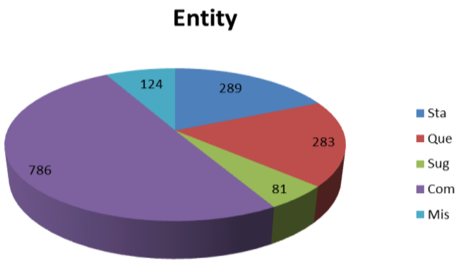

Figure 3. Speech Act Distribution (Entity)

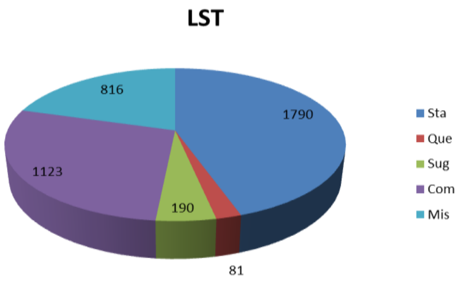

Figure 4. Speech Act Distribution (LST)

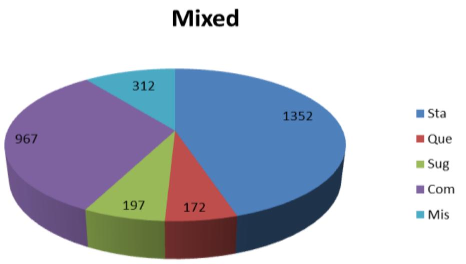

Figure 5. Speech Act Distribution (Mixed)

For each category, we use two labeled/unlabeled data settings, with labeled data accounting for 5% and 10% of the total so that the labeled/unlabeled ratios are set at approximately 1:19 and 1:9. The labeled data in each category are randomly selected in a stratified way: using the same percentage to select labeled data with each speech act type. The stratified selection is intended to keep the speech act distributions in both labeled and unlabeled data. Table 2 and Table 3 list the details of data splitting using the two settings.

Category	# Labeled	# Unlabeled	Total
News	155	2995	3150
Entity	72	1391	1463
LST	198	3802	4000
Mixed	147	2853	3000

Table 2. Stratified Data Splitting with 5% as Labeled

[1] http://www4.comp.polyu.edu.hk/~csrzhang

Category	# Labeled	# Unlabeled	Total
News	312	2838	3150
Entity	144	1319	1463
LST	399	3601	4000
Mixed	298	2702	3000

Table 3. Stratified Data Splitting with 10% as Labeled

For comparison with supervised learning, we also use inductive SVM. The inductive and transductive SVM classifications are implemented by using the SVM^{light} tool[2] with a linear kernel. For the graph-based label propagation method, we populate the similarity matrix **W** with weights calculated by a Gaussian function. Given two tweets $\mathbf{x}_i$ and $\mathbf{x}_j$,

$$w_{ij} = \exp(-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2}{2\sigma^2})$$

where $\|.\|$ is the L2 norm. Empirically, the Gaussian function measure leads to better results than other measures such as cosine. Then we convert the graph to an *εNN* graph (Zhu and Goldberg 2009) by removing edges with weight less than a threshold because the *εNN* graph empirically outperforms the fully connected graph. The threshold is set to be $\mu + \sigma$, the mean of all weights plus one standard deviation.

5.2 Results

To better evaluate the performance of semi-supervised learning on speech act recognition in Twitter, we report the classification scores for both multi-class and individual classes, as well as confusion matrices.

Multi-class Evaluation

Table 4 lists the macro-average F scores and weighted average F scores for all classifiers and all categories at the 5% labeled data setting. Macro-average F is chosen because it gives equal weight to all classes. Since some classes (e.g., sta) have much more instances than others (e.g., que), macro-average F ensures that significant score change on minority classes will not be overshadowed by small score change on majority classes. In contrast, weighted average F is calculated according to class instance numbers, which is chosen mainly because we want to compare the result with supervised learning (reported in Zhang et al. 2011 and Figure 1). In this and the following tables, iSVM, tSVM, and GLP denote inductive SVM, transductive SVM, and graph-based label propagation.

	Macro-average F			**Weighted average F**		
	iSVM	*tSVM*	*GLP*	*iSVM*	*tSVM*	*GLP*
News	.374	.502	.285	.702	.759	.643
Entity	.312	.395	.329	.493	.534	.436
LST	.295	.360	.216	.433	.501	.376
Mixed	.383	.424	.245	.539	.537	.391

Table 4. Multi-class F scores (5% labeled data)

Almost without exception, transductive SVM achieves the best performance. Measured by macro-average F, it outperforms inductive SVM with a gain of 10.7% (Mixed) to 34.2% (News). Consistent with supervised learning results, semi-supervised learning results degrade with News > Entity > LST, indicating that both semi-supervised learning and supervised learning are sensitive to dataset characteristics. More uniform tweet set (e.g., News) leads to better classification and greater improvement by semi-supervised learning. That also explains why the Mixed category, composed of the most diversified tweets, benefits least from semi-supervised learning.

Conversely, supervised learning (inductive SVM) on the Mixed category benefits from the data hodgepodge even though the test data are 19 times the training data. Its macro-average F is higher than the other categories although it does not have the most training data. Its weighted-average F using inductive SVM is even higher than using transductive SVM.

It is a little surprising to find that the graph-based label propagation performs very poorly. In all but one place, the GLP score is lower than its iSVM counterpart. This may indicate that the graph method cannot adapt well to the multi-class scenario and we will show more evidences in the next two sections.

To understand the effectiveness of semi-supervised learning, a better way than doing numerical calculation is juxtaposing semi-supervised data settings with their comparable supervised data settings, which is shown in Table 5. The supervised data settings are of those with the closest weighted average F (waF) to the semi-supervised (tSVM) waF from our previous results (Figure 1).

[2] http://svmlight.joachims.org/

	# labeled	*labeled :unlabeled*	*waF*
	Semi-supervised (tSVM)		
News	155	1 : 19	.759
Entity	72	1 : 19	.534
LST	198	1 : 19	.501
Mixed	147	1 : 19	.537
	Supervised (with closest waF)		
News	945	1 : 0.3	.768
Entity	146	1 : 1	.589
LST	800	1 : 0.5	.501
Mixed	861	1 : 1	.596

Table 5. Semi-supervised Learning vs. Supervised Learning

Obviously semi-supervised learning by transductive SVM can achieve classification performance comparable to supervised learning by inductive SVM, with less training data and much lower labeled/unlabeled ratio. This shows that semi-supervised learning such as transductive SVM holds much promise for scalable speech act recognition in Twitter.

It is tempting to think that with more labeled data and higher labeled/unlabeled ratio, semi-supervised learning performance should improve. To put this conjecture to test, we double the labeled data (from 5% to 10%) and labeled/unlabeled ratio (from 1/19 to 1/9), with results in Table 6.

	Macro-average F			**Weighted average F**		
	iSVM	*tSVM*	*GLP*	*iSVM*	*tSVM*	*GLP*
News	.403	.524	.298	.731	.762	.647
Entity	.441	.440	.311	.587	.575	.406
LST	.335	.397	.216	.459	.512	.384
Mixed	.435	.463	.284	.557	.553	.415

Table 6. Multi-class F scores (10% labeled data)

Compared with Table 4, increased labeled data does lead to some improvement, but not much as we would expect, the largest gain being 15.9% (macro-average F on Mixed, using GLP). Note that this is achieved at the cost of labeling twice as much data and predicting half as much. In contrast, the inductive SVM performance is improved by as much as 41.3% (macro-average F on Entity). Such evidence shows that semi-supervised learning of speech acts in Twitter benefits disproportionately little from increased labeled data, or at least the gain is not worth the pain. In fact, this is good news for scalable speech act recognition.

Individual Class Evaluation

For more microscopic inspection, we also report the classification results on individual classes for all categories. In Table 7, we list the rankings of F measures by each classifier for each speech act type and each category. The one-letter notations *i*, *t*, *g* are short for iSVM, tSVM, and GLP. Therefore, *t* > *g* > *i* means tSVM outperforms GLP, which outperforms iSVM, in terms of F measure. The labeled data are 5%.

	Sta	**Que**	**Sug**	**Com**	**Mis**
News	*t >g>i*	*t >i>g*	*t >i>g*	*t >i>g*	*t >g>i*
Entity	*t >g>i*	*t >i>g*	*g >t>i*	*i >t>g*	*t >g>i*
LST	*i >g>t*	*t >i>g*	*i >t>g*	*t >i>g*	*t >g>i*
Mixed	*i >t>g*	*t >i>g*	*t >i>g*	*i >t>g*	*t >g>i*

Table 7. Classifier Rankings for Each Speech Act Type and Category (5% Labeled Data)

In 15 out of the 20 rankings, transductive SVM or graph-based label propagation beats inductive SVM, which shows the efficacy of semi-supervised learning in this class-based perspective. Transductive SVM is the champion, claiming 14 top places.

We also find that the overall performance of graph-based label propagation is the poorest, claiming 12 out of 20 bottom places. After inspecting the data, we observe that the underlying assumption of GLP that similar objects belong to the same class is questionable for speech act recognition in Twitter. Tweets with different speech acts (e.g., question and comment) may appear very similar on the graph. The maximal margin approach is apparently more appropriate for our problem.

On the other hand, the GLP performance evaluated on individual classes is better than evaluated on the multi-class if we compare Table 7 and Table 4, where GLP is almost always the lowest achiever. This indicates that in multi-class classification, GLP suffers further from the one-vs-all converting scheme, a point we will make clearer in the following.

Confusion matrices

Confusion matrix provides another perspective to understand the multi-class classification performance. For brevity's sake, we present the confusion matrices of the three classifiers on the News category with 5% labeled data in Figure 6 to Figure 8. Similar patterns are also observed for the other categories and with 10% labeled data. Note that the rows represent true classes and the columns represent predicted classes.

	Sta	**Que**	**Sug**	**Com**	**Mis**
Sta	2043	0	5	14	0
Que	46	7	2	9	0
Sug	211	1	61	21	0
Com	276	2	10	164	0
Mis	120	0	1	2	0

Figure 6. Confusion Matrix of iSVM (News, 5% Labeled Data)

	Sta	**Que**	**Sug**	**Com**	**Mis**
Sta	1848	4	56	90	64
Que	19	17	7	20	1
Sug	95	0	158	31	10
Com	143	5	19	275	10
Mis	94	3	4	15	7

Figure 7. Confusion Matrix of tSVM (News, 5% Labeled Data)

	Sta	**Que**	**Sug**	**Com**	**Mis**
Sta	1852	0	4	11	195
Que	19	6	0	0	39
Sug	123	0	25	2	144
Com	134	0	0	47	271
Mis	102	0	0	1	20

Figure 8. Confusion Matrix of GLP (News, 5% Labeled Data)

The News category is typically biased towards the statement speech act, which accounts for 69% of the total tweets according to Figure 2. As a result, the iSVM tends to classify tweets of the other speech acts as statement. Figure 6 also shows that the prediction accuracy is correlated with the training amount. The two classes with the least training data, question and miscellaneous, demonstrate the lowest accuracy. Clearly, supervised learning suffers from training data deficiency.

Both tSVM and GLP show the effect of leveraging unlabeled data as they assign new labels to some instances wrongly classified as statement. Transductive SVM is more successful in that it moves most of the Sug and Com instances to the diagonal. The situation for Que and Mis is also better, though the prediction accuracy still suffers from lack of training data. Figure 8, however, reveals an intrinsic problem of applying graph-based label propagation to multi-class classification. Most instances are predicted as either Sta or Mis. The wrong prediction as Mis cannot be explained by imbalance of training data. Rather, it is due to the fact that the single-class scores for Mis after smoothing on the graph are generally higher than those for Que, Sug, or Com. In other words, the graph-based method is highly sensitive to class differences when multi-class prediction is converted from single-class predictions on a scheme like one-vs-all.

In contrast, transductive SVM does not suffer much from class differences according to Figure 7, proving to be more suitable for multi-class classification than graph-based label propagation.

5.3 Summary

For the task of recognizing speech acts in Twitter, we have made some interesting findings from the extensive empirical study. To wrap up, let's summarize the most important of them in the following.

1) Semi-supervised learning approaches, especially transductive SVM, perform comparably to supervised learning approaches, such as inductive SVM, with considerably less training data and lower training/test ratio. Increasing training data cannot improve performance proportionately.

2) Transductive SVM proves to be more effective than graph-based label propagation for our task. The performance of the latter is hurt by two factors: a) the inappropriate assumption about similar tweets having the same speech act and b) its vulnerability to class differences under the one-vs-all multi-class conversion scheme.

3) For supervised learning as well as semi-supervised learning for multi-class classification, training data imbalance poses no lesser threat than training data deficiency.

6. Conclusion and Future Work

Speech act recognition in Twitter facilitates content-based user behavior study. Realizing that it is obsessed with insufficient training data, we start where previous research left.

We are not aware of previous study of semi-supervised learning of speech acts in Twitter and in this paper we contribute to scalable speech act recognition by drawing conclusions from extensive experiments. Specifically, we

1) extend the work of (Zhang et al. 2011) by establishing the practicality of semi-supervised learning that leverages the knowledge of unlabeled data as a promising solution to insufficient training data;

2) show that transductive SVM is more effective than graph-based label propagation for our problem, which aptly extends the maximal margin approach to unlabeled data and is more amenable to the multi-class scenario;

3) provide detailed empirical evidences of multi-class and single-class results, which can inform future extensions in this direction and design of practical systems.

At this stage, we are not sure whether the one-vs-all scheme is a bottleneck to one class-oriented classifiers (it appears to be so for the graph-based method). Therefore we will next explore other multi-class conversion schemes and also consider semi-supervised learning using inherently multi-class classifiers such as Naïve Bayes or Decision Tree. In the future, we will also explore unsupervised approaches to recognizing speech acts in Twitter.

Acknowledgments

The work described in this paper was supported by the grants GRF PolyU 5217/07E and PolyU 5230/08E.

References

Austin, J. 1962. *How to Do Things with Words*. Oxford: Oxford University Press.

Cohen, W., Carvalho, V., and Mitchell, T. 2004. Learning to Classify Email into “Speech Acts”. In *Proceedings of Empirical Methods in Natural Language Processing* (*EMNLP-04*), 309–316.

Crystal, D. 2006. *Language and the Internet, 2nd edition*. Cambridge, UK: Cambridge University Press.

Crystal, D. 2011. *Internet linguistics*. London: Routledge.

Culp M. and Michailidis, G. 2007. An Iterative Algorithm for Extending Learners to a Semisupervised Setting. In *The 2007 Joint Statistical Meetings (JSM).*

Deshpande S. S., Palshikar, G. K., and Athiappan, G. 2010. An Unsupervised Approach to Sentence Classification, In *International Conference on Management of Data* (*COMAD 2010*), Nagpur, India.

Dhillon, R., Bhagat, S., Carvey, H., and Shriberg, E. 2004. *Meeting Recorder Project: Dialog Act Labeling Guide*. Technical report, International Computer Science Institute.

Erkan, G., Özgür, A., and Radev, D. 2007. Semi-Supervised Classification for Extracting Protein Interaction Sentences Using Dependency Parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 228–237.

Feng, D., Shaw, E., Kim, J., and Hovy. E. H. 2006. Learning to Detect Conversation Focus of Threaded Discussions. In *Proceedings of HLT-NAACL*, 208–215.

Haffari G.R. and Sarkar. A. 2007. Analysis of semi-supervised learning with the Yarowsky algorithm. In *23rd Conference on Uncertainty in Artificial Intelligence (UAI).*

Jeong, M., Lin, C-Y., and Lee, G. 2009. Semi-supervised Speech Act Recognition in Emails and Forums. In *Proceedings of EMNLP*, pages 1250–1259.

Joachims, T. 1999. Transductive Inference for Text Classification using Support Vector Machines. In *Proceedings of the 16th International Conference on Machine Learning (ICML)*.

Jurafsky, D., Shriberg, E., and Biasca, D. 1997. *Switchboard SWBD-DAMSL Labeling Project Coder’s Manual, Draft 13*. Technical report, University of Colorado Institute of Cognitive Science.

Medlock, B., and Briscoe, T. 2007. Weakly Supervised Learning for Hedge Classification in Scientific Literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 992–999.

Searle, J. 1975. Indirect speech acts. In P. Cole and J. Morgan (eds.), *Syntax and semantics, vol. iii: Speech acts* (pp. 59–82). New York: Academic Press.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R. Van Ess-Dykema, C., and Meteer, M. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339–373.

Vapnik, V. 1998. *Statistical Learning Theory*. New York: John Wiley & Sons.

Yarowsky, D. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-1995)*, 189–196.

Zhang, R., Gao, D., and Li, W. 2011. What Are Tweeters Doing: Recognizing Speech Acts in Twitter. In *AAAI-11 Workshop on Analyzing Microtext.*

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., and Scholkopf, B. 2004. Learning with Local and Global Consistency. *Advances in Neural Information Processing Systems (NIPS), vol. 16*, Cambridge, MA: MIT Press.

Zhu, X., Ghahramani, Z., and Lafferty, J. D. 2003. Semi-supervised Learning Using Gaussian Fields and Harmonic Functions. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, 912–919, Washington, DC.

Zhu, X. and Goldberg, A. B., 2009. *Introduction to Semi-Supervised Learning*. Morgan & Claypool Publishers.