Semi-supervised Classification of Twitter Messages for Organization Name Disambiguation

Shu Zhang[1], Jianwei Wu[2], Dequan Zheng[2], Yao Meng[1] and Hao Yu[1]
[1] Fujitsu Research and Development Center, Beijing, China
{zhangshu, mengyao, yu}@cn.fujitsu.com
[2] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
{jwwu, dqzheng}@mtlab.hit.edu.cn

Abstract

In this paper, we probe the problem of organization name disambiguation on twitter messages. This task is challenging due to the fact of lacking sufficient information in a tweet message. Instead of conventional methods based on mining external information from web sources to enrich information about organization, we propose to mine the relationship among tweets in data set to utilize context information for disambiguation. With a small scale of labeled tweets, we propose LP-based and TSVM-based semi-supervised methods to classify tweets. We aim to mine both related and non-related information for a given organization. The experiments on WePS-3 show that proposed methods are effective.

1 Introduction

Twitter is an online social networking and microblogging service, which rapidly gained worldwide popularity. How to retrieval, analyze and monitor Twitter information has been receiving a lot of attention in natural language processing and information retrieval research community (Kwak, *et al.*, 2010; Boyd, *et al.*,2010; Tsagkias, *et al.*, 2011). One of the essential things of these researches is first to get the information which is related to the studied entity. This is caused by the ambiguity of entities. For example, the name of company "*Apple*" has a separate meaning referring to one kind of fruit. The word "*Amazon*" also could refer to river or company.

In this paper, we focus on finding related tweets for a given organization, which can be treated as a binary classification problem. Assuming that tweets are retrieved by a query, such as "*apple*", the task is to classify whether each retrieved tweet is relevant to the target organization ("*Apple Inc.*") or not. However, constructing such a classifier is a challenging task, as tweets are short and informal. Additionally, the information about a given organization is limited, which is difficult to cover the word occurrences in the given organization related tweets.

Different from previous work on mining external information from web sources to enrich information about the given organization, we propose to mine the relationship among retrieved tweets in data set. With a small scale of labeled tweets, we propose semi-supervised methods to mine the relationships between labeled and unlabeled tweets for the given organization.

The remainder of the paper is organized as follows: Section 2 describes the related work on name disambiguation. Section 3 gives problem description and an overview of our approach. Section 4 and Section 5 present LP-based and TSVM-based semi-supervised methods to classify tweets. Section 6 gives the experiments and results. Finally section 7 summarizes this paper.

2 Related Work

Twitter contains little information in each tweet, with no more than 140 characters. This makes the tasks of analyzing Twitter messages more challenge, and attracts much interest from the research community in recent years (Meij *et al.*, 2012; Liu *et al.*, 2011; Sriram *et al.*, 2011).

The most related works are WePS-3 Online Reputation Management[1] held in 2010, which aims to identify tweets which are related to a given company (Amigó *et al.*, 2010).

In WePS-3, the research of (Yerva *et al.*, 2010) shows the best performance in the evaluation

[1] http://nlp.uned.es/weps/

International Joint Conference on Natural Language Processing, pages 869–873,
Nagoya, Japan, 14-18 October 2013.

campaign. They adopt SVM classifier with external resources, including Wordnet, metadata profile, category profile, Google set, and user feedback, to enrich the information of the given organization. Yoshida *et al.* (2010) classify organization names into "organization-like names" or "general-word-like names". Kalmar (2010) adopts bootstrapping method to classify the tweets. The research of (García-Cumbreras *et al.*, 2010) shows the named entities in tweets are appropriate for certain company names.

There are some similar works. Perez-Tellez *et al.* (2011) adopt clustering technique to solve the problem of organization name disambiguation. Focus on identifying relevant tweets for social TV, Dan *et al.* (2011) propose a bootstrapping algorithm utilizing a small manually labeled dataset, and a large dataset of unlabeled messages.

Different from their works, we utilize semi-supervised methods to classify the tweets. We aim to transfer related or unrelated information of the given organization among tweets based on a small scale of labeled data.

Compared with bootstrapping algorithm, which is based on a local consistency assumption, LP algorithm is based on a global consistency assumption, and can effectively capture the natural clustering structure in both the labeled and unlabeled data to smooth the labeling function.

3 Overview

3.1 Problem Statement

Given a set of tweets and an organization name, the goal is to decide whether each tweet in the set talks about the given organization or not.

In detail, the input information per tweet contains: the tweet identifier, the entity name, the query used to retrieve the tweet, the author identifier and the tweet content.

For each organization in the dataset, it gives the organization name and its homepage URL.

The output per tweet is True or False tag corresponding to related or non-related with the given organization.

3.2 Our Method

In this paper, we propose semi-supervised methods to classify tweets for a given organization. This is considered from the following two points:

- Organization information automatically mined from web pages is limited, which could not cover the potentially infinite words occurred in tweets. However, how to mine the high quality organization related information is also a problem.
- Both positive and negative samples are important for classification task. Though, it is possible to mine organization related information as positive sample from web by some key words or human input. However, it is difficult to obtain negative information about the other meanings of the given organization name which do not refer to the given organization.

Therefore, instead of mining external information from web sources to enrich information about organization, we propose to mine information directly from tweet set. The organization related information is extracted from the positive samples, which reflects keywords related to the given organization in tweets. The information extracted from the negative samples, gives the possible different interpretations of the given organization name.

With a small scale of labeled tweets for a given organization name, we utilize LP and TSVM based semi-supervised classifiers to mine unlabeled tweets, which will be described in the following section in detail.

4 LP Based Semi-supervised Classifier

Label Propagation (LP) is a graph-based semi-supervised algorithm, proposed by Zhu *et al.* (2002). The main idea of graph-based semi-supervised learning is to use pair-wise similarities between instances to enhance classification accuracy. It is a diffusion process on graphs, where the information is propagated from the labeled instances to the rest of unlabeled instances.

LP algorithm is to represent labeled (served as seeds) and unlabeled examples as nodes in a connected graph, then propagating the label information from any vertex to nearby nodes through weighted edges iteratively, finally get the labels of unlabeled examples after the propagation process converges. The labels of unlabeled examples are determined by considering both the similarity between labeled and unlabeled examples, and the similarity between unlabeled examples (Chen, *et al.*,2008).

LP algorithm has achieved good performance in many applications, such as noun phrase anaphoricity in coreference resolution (Zhou, *et al.*, 2009), word sense disambiguation (Niu, *et al.*, 2005) and entity relation extraction (Chen, *et al.*, 2006).

4.1 Graph Building

Let $X = \{x_i\}_{i=1}^{n}$ be a set of tweets for a given organization, where x_i represents *ith* tweet, *n* is the total number of tweets. The first *l* tweets are labeled $(x_1, y_1)...(x_l, y_l)$, $Y_L = \{y_1, ..., y_l\}$ are labels. Here, $y_i \in C$, C refers to two known classes (*True* or *False*) for this task. The others $(x_{l+1}, y_{l+1}) \ldots (x_{l+u}, y_{l+u})$ are the unlabeled tweets, where $Y_U = \{y_{l+1}, ..., y_{l+u}\}$ are unknown.

For the graph, the nodes represent both labeled and unlabeled tweets. The edge between any two nodes x_i and x_j is weighted by some distance measure. Based on assumption, the closer the two nodes are in some distance measure, the larger the weight w_{ij}, which is defined as follows:

$$w_{ij} = \exp(-\frac{s_{ij}^2}{\sigma^2}) = \exp(-\frac{Cos^2(x_i, x_j)}{\sigma^2})$$

Where s_{ij} is the distance measure, we adopt cosine similarity to measure two nodes x_i and x_j. σ is a constant parameter to scale the weights.

For measuring the similarity of two nodes, we adopt two types of features to represent each tweet: one is the unigram word unit, the other is 4-gram character unit.

Unigram word: the words contain in a tweet after filtering stop words.

4-gram character unit: the possible 4-gram character for each unigram word.

The tweet is short and informal. There are little information contain in one tweet. One keyword missing may lead the change of the tweet's classification result. Therefore, we adopt character unit as feature to allow the mistake of spelling in some extent.

4.2 Algorithm

All nodes in graph have soft labels that can be interpreted as distribution over labels. The label of a node is propagated to all nodes through the edges. Larger edge weights allow labels to travel through easier. Define a $n \times n$ probabilistic transition matrix *T*, ($n = l + u$).

$$T_{ij} = P(j \to i) = \frac{w_{ij}}{\sum_{k=1}^{l+u} w_{kj}}$$

Here T_{ij} is the probability to jump from node *j* to node *i*. We define a $(l+u) \times C$ label matrix *Y*, the *i*th row representing the label probability distribution of node x_i.

The label propagation algorithm is as follows:

(1) Propagate $Y \leftarrow TY$
(2) Row-normalize *Y*, to maintain the label probability interpretation
(3) Clamp the labeled tweets, replace the Y_L with the initial value
(4) Repeat from step (1) to (3) until *Y* converges

Here, we make use of JUNTO Label Propagation toolkit[2] to implement this algorithm.

5 Transductive SVM

Transductive Support Vector Machines (TSVM) is a semi-supervised learning method, which can be treated as an extension of SVM by introducing unlabeled data. Similar with SVM, TSVM tries to label the unlabeled data, and find the maximum margin separation hypersurface that separates the positive and negative instances of labeled data and the unlabeled data. The basic idea of TSVM is to seek a decision surface away from the dense regions of unlabeled data.

For the given labeled tweets $\{(x_i, y_i) \mid x_i \in R^n$ $y_i \in \{-1, +1\}\}_{i=1}^{L}$, y_i refers to two known classes (*True* or *False*) for this task, and unlabeled data $\{x_j^* \mid x_j^* \in R^n\}_{j=1}^{L^*}$. This can be written as minimizing

$$\arg\min_{w,b,\xi,y^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{L}\xi_i + C^*\sum_{j=1}^{L^*}\xi_j^*$$

Subject to

$$\forall i : y_i w \bullet \phi(x_i) + b \geq 1 - \xi_i$$
$$\forall j : y_j^* w \bullet \phi(x_j) + b \geq 1 - \xi_j^*$$
$$\forall i : \xi_i \geq 0$$
$$\forall j : \xi_j^* \geq 0$$
$$\forall j : y_j^* \in \{-1, +1\}$$
$$w \in R^m, b \in R$$

Similar with LP, we adopt two types of features to represent each tweet: one is the unigram word unit, the other is 4-gram character unit. Here, we make use of SVMLight tools to implement this algorithm.

6 Experiments and Results

6.1 Corpus and Evaluation Metric

We have conducted experiments on the WePS-3 task 2 data. The test data contain about 50 organization names with about 450 tweets for each organization.

[2] http://code.google.com/p/junto/

	P+	*R*+	*F*+	*P*-	*R*-	*F*-
LP	0.8097	0.5008	0.4120	0.8059	0.5593	0.5166
TSVM	0.6683	0.6969	**0.7144**	0.6942	0.6484	**0.6972**
Top_1	0.7108	0.7445	0.6264	0.8443	0.5195	0.5606
Top_2	0.7546	0.5409	0.4935	0.7413	0.6049	0.5651
Top_3	0.7410	0.6157	0.5062	0.7365	0.4911	0.4683
Baseline (NR)	1.0000	0.0000	0.0000	0.5652	1.0000	0.6563
Baseline (R)	0.4348	1.0000	0.5274	1.0000	0.0000	0.0000

Table 1. Performances of semi-supervised methods and other systems

The task is to classify the tweets related or non-related with the given organization, it belongs to classification task. Therefore, we measure the performance by *accuracy*, *precision*, *recall* and *F-measure*.

6.2 Results and Analysis

Based on the test data, we testify the performance of our proposed methods.

Seed selection for semi-supervised classifiers

We random select 100 tweets as seeds from the test data for each organization name, which is about 20% for tweet set.

Decreasing the influence of seed selection for the performances of semi-supervised classifiers, we try out the experiments five times and get the average values for the final results.

Performance of semi-supervised classifiers

For comparison, we select five system results as references, three of them are the top 3 systems in WePS contest, the other two systems are the baseline systems. Two baseline systems tag all tweets as related (Baseline (R)) or non-related (Baseline (NR)) to each organization.

Figure 1 and Table 1 show the performances of semi-supervised methods and other systems.

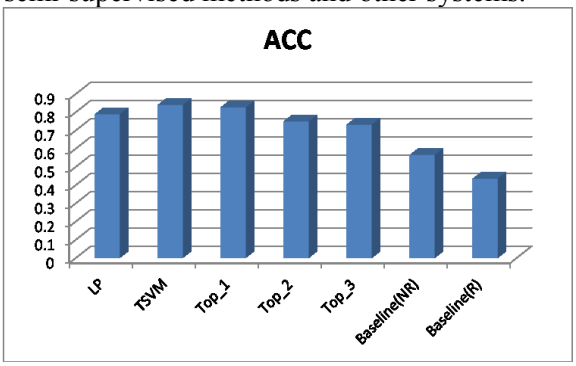

Figure 1. Accuracies of semi-supervised methods and other systems

In Figure 1, the *accuracy* of Baseline (NR) is higher than that of Baseline (R), which shows there are more unrelated tweets in the whole test data, the disambiguation of tweets is necessary. The accuracies of our proposed methods and Top 3 systems are all much higher than those of two baselines. It proves that adopting some methods to disambiguate tweets is feasible.

The accuracies of our proposed semi-supervised methods based on LP and TSVM are both higher than that of Top_2 system. The accuracy of TSVM is 0.8391, which is higher than that of Top_1 (0.8267). It proves that semi-supervised methods are effective for this task. Instead of mining web sources, it is also effective to mine the information among tweets, especially which including both related and non-related information about the organization name.

In Table 1, it shows the performance of each system on *precision*, *recall* and *F-measure*. The values are calculated on the average performance for each organization name in test data set. Though *P*+ and *R*+ values of TSVM are not the highest ones, the *F*+ value is highest among the five systems. *F*- value is also the highest one. it shows that TSVM-based classifier gets gets the best balance between precision and recall for classification. *F*+ value is important to measure the ability of finding the related tweets to a given organization.

7 Conclusion

In this paper, we probe the problem of organization name disambiguation on twitter information. We utilize LP and TSVM based semi-supervised method to implement the disambiguation system. The experiments on WePS-3 show that both LP-based classifier and TSVM-based classifier are effective. Especially, TSVM-based classifier gets higher performance than that of the best result in WePS contest, which proves that semi-supervised method is a feasible way to classify the related tweets information for a given organization on Twitter.

References

Enrique Amigó, Javier Artiles, Julio Gonzalo, Damiano Spina, Bing Liu, and Adolfo Corujo. 2010. WePS-3 Evaluation Campaign: Overview of the Online Reputation Management Task. In Proceedings of 3rd Web People Search Evaluation Workshop.

Surender R. Yerva, Zoltán Miklós, and Karl Aberer. 2010. It was Easy, when Apples and Blackberries were only Fruits. In Proceedings of 3rd Web People Search Evaluation Workshop.

Minoru Yoshida, Shin Matsushima, Shingo Ono, Issei Sato, and Hiroshi Nakagawa. 2010. ITC-UT: Tweet Categorization by Query Categorization for On-line Reputation Management. In Proceedings of 3rd Web People Search Evaluation Workshop.

Paul Kalmar. 2010. Bootstrapping Websites for Classification of Organization Names on Twitter. In Proceedings of 3rd Web People Search Evaluation Workshop.

M.A. García-Cumbreras, M. García-Vega M, F. Martínez-Santiago and J.M. Peréa-Ortega. 2010. SINAI at WePS-3: Online Reputation Management. In Proceedings of 3rd Web People Search Evaluation Workshop.

Fernando Perez-Tellez, David Pinto, John Cardiff, and Paolo Rosso. 2011. On the Difficulty of Clustering Microblog Texts for Online Reputation Management. In Proceedings of 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT.

Ovidiu Dan, Junlan Feng, and Brian D. Davison. 2011. A Bootstrapping Approach to Identifying Relevant Tweets for Social TV. In Proceedings of 5th International AAAI Conference Weblogs and Social Media.

Xuan Hieu Phan, Le Minh Nguyen, and Susumu Horiguchi. 2008. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. In Proceeding of 17th WWW, pages 91-100.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a Social Network or a News Media? In Proceeding of 19th WWW, pages 591-600.

Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In Hawaii International Conference on System Sciences, pages 1-10.

Manos Tsagkias, Maarten de Rijke, and Wouter Weerkamp. 2011. Linking Online News and Social Media. In Proceedings of 4th ACM Web Search and Data Mining, pages 565-574.

Edgar Meij, Wouter Weerkamp, and Maarten de Rijke. 2012. Adding Semantic to Microblog Posts. In proceedings of 5th ACM Web Search and Data Mining, pages 563-572.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing Named Entities in Tweets. In Proceedings of 49th Annual Meeting of the Association for Computational Linguistics, pages 359-367.

Bharath Sriram, David Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2011. Short Text Classification in Twitter to Improve Information Filtering. In Proceedings of the ACM SIGIR 2011, pages 841-842.

Guo Dong Zhou and Fang Kong. 2009. Global Learning of Noun Phrase Anaphoricity in Coreference Resolution via Label Propagation. In Proceedings of Empirical Methods in Natural Language Processing, pages 978-986.

Zheng Yu Niu, Dong Hong Ji, and Chew Lim Tan. 2005. Word Sense Disambiguation Using Label Propagation Based Semi-Supervised Learning. In Proceedings of 43rd Annual Meeting on Association for Computational Linguistics, pages 395-402.

Jin Xiu Chen, Dong Hong Ji, Chew Lim Tan, and Zheng Yu Niu. 2006. Relation Extraction Using Label Propagation Based Semi-supervised Learning. In Proceedings of 21st International Conference on Computational Linguistics and 44th Annual Meeting on Association for Computational Linguistics, pages 129-136.

Jin Xiu Chen, and Dong Hong Ji. 2008. Graph-Based Semi-supervised Relation Extraction. In Journal of Software, 19(11): 2843-2852.

Xiao Jin Zhu and Zou Bin Ghahramani. 2002. Learning from Labeled and Unlabeled Data with Label Propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University