Supervised and Semi-supervised Methods based Organization Name Disambiguity

Shu Zhang[a], Jianwei Wu[b], Dequan Zheng[b], Yao Meng[a], Yingju Xia[a], and Hao Yu[a]

[a]Fujitsu Research and Development Center
Dong Si Huan Zhong Rd, Chaoyang District, Beijing and 0086, China
{zhangshu, mengyao, yjxia, yu}@cn.fujitsu.com

[b]School of Computer Science and Technology, Harbin Institute of Technology
No.92, Xidazhi Street, Harbin 150001, China
{jwwu, dqzheng}@mtlab.hit.edu.cn

Abstract. Twitter is a widespread social media, which rapidly gained worldwide popularity. Pursuing on the problem of finding related tweets to a given organization, we propose supervised and semi-supervised based methods. This is a challenging task due to the potential organization name ambiguity. The tweets and organization contain little information. The organizations in training data are different with those in test data, which leads that we could not train a classifier to a certain organization. Therefore, we induce external resources to enrich the information of organization. Supervised and semi-supervised methods are adopted in two stages to classify the tweets. This is a try to utilize both training and test data for this specific task. Our experimental results on WePS-3 are primary and encouraging, they prove the proposed techniques are effective in performing the task.

Keywords: Twitter, name ambiguity, online reputation management.

1 Introduction

Twitter is an online social networking and microblogging service, which rapidly gained worldwide popularity, with 200 million users as of 2011[1], generating over 200 million tweets and handling over 1.6 billion search queries per day[2]. How to manage this information to grasp the response of people to governmental policies, the feedback and comment of people on commercial products have received considerable attention in research community. There are some researches such as opinion mining, online reputation management, which focus on monitoring user generated media. One of the essential things of these researches is first to get the information which is related to the studied entity, such as product, company, or certain event.

This paper focuses on finding related tweets to a given organization. This is a challenging task due to the potential organization name ambiguity. For example, the name of company "*Apple*" which has a separate meaning fruit apple. The word "*Amazon*" could be used to refer river or company. Filtering spurious name matches is important to effectively detect and analyze relevant contents that people say about the organization.

To overcome the problem that the tweets and organizations contain little information, we induce external resources to enrich the information of organization. The organizations in training data are different with those in test data, which leads that we could not train a classifier to a certain organization. Therefore, supervised and semi-supervised methods are adopted in

[1] http://www.bbc.co.uk/news/business-12889048
[2] http://blog.twitter.com/2011/08/your-world-more-connected.html

two stages to classify the tweets. This is a try to utilize both training and test data for this specific task.

The remainder of the paper is organized as follows: Section 2 describes the related work on name disambiguity. Section 3 gives overview of the problem and our methods. Section 4 presents supervised method to classify tweets. Section 5 introduces semi-supervised method to classify the tweets which is a step of modifying initial classification results gotten by supervised method. Section 6 gives the experiments and results. Finally section 7 summarizes this paper.

2 Related work

In recent years, online social networks such as Twitter have attracted much interest from the research community. Twitter differs from the traditional user generated media. It allows users to generate each message with no more than 140 characters. Little context information is available. Therefore, it is a challenge for monitoring and analyzing them.

Dan *et al.* (2011) focus on identifying relevant tweets for social TV, they propose a bootstrapping algorithm which uses a small manually labeled dataset, a large dataset of unlabeled messages, and some domain knowledge to derive a classifier to filter microblogging messages which discuss television show. They extract features which contain general terms commonly associated with watching TV.

WePS-3 Online Reputation Management[3] held in 2010, aimed to identify tweets which are related to a given company. It provides standard training and test dataset that enable researchers to carry out and evaluate their methods (Amigó *et al.*, 2010). For this task, the research of (Yerva *et al.*, 2010) shows the best performance in the evaluation campaign. They adopt support vector machines (SVM) classifier with external resources, which includes Wordnet, metadata profile, category profile, Google set, and user feedback. To overcome the problem of tweets containing little context information, they create several profiles with external resources as a model for each company. Yoshida *et al.* (2010) classify organization names into "organization-like names" or "general-word-like names". They categorize each query in the first stage, and categorize each tweet in the second stage using the rules customized for each class of queries. 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. Tsagkias *et al.* (2010) prove that a general classifier can be employed to predict the presence of any company in Twitter.

Perez-Tellez *et al.* (2011) propose term expansion to the ambiguous words and words which highly co-occur with it.

Our work is different from theirs: supervised and semi-supervised methods are utilized in different stages for the classification of tweets. For the task, the set of organization names in the training and test corpora are different. The model could not be trained for a certain organization. Therefore, our method aims to utilize both training data and the test data to improve the accuracy of the performance. In detail, the tweets are firstly classified by Maximum Entropy classifier trained on training data. Then Label Propagation Algorithm is utilized to classify tweets based on the test data.

3 Overview

3.1 Problem Statement

Given a set of tweets and an organization name, the task is to judge that each tweet in the set is related to the organization or not.

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

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.

Compared with the traditional classification task, this task has some challenges. One is tweets and organization name contain little information, context information is limited. The other is the set of organization names in the training and test data are different. That means no organization is present in both training data and test data, the features or model built on training data is not very suitable for the test data.

3.2 Our Method

Overcome the challenges of this task, we import external resources to get more information about the organization, such as related homepage, related Wikipedia page, and GoogleSet. We propose to combine supervised and semi-supervised methods to classify the tweets, aims to utilize the training data and at the same time dig out the proper information in the test data. The system includes the following parts:

(1) Features extraction and representation. It includes both the features from the tweets and the features of organization from external resources.

(2) Maximum entropy classifier. It is trained with the features extracted in Step (1) on the training data.

(3) Label Propagation. It includes two parts: seeds selection and graph construction. We select seeds based on the result gotten in Step (2). For one certain organization, we construct graph based on the relationship between each tweets on test data.

(4) Rule-based modification. Here, we only give one rule which processes the organization name containing more than one word, such as “Yale University”. We think that “Yale University” contains more semantic information to distinguish it from other entity. The rule is if the tweet contains the full entity name (more than one word) then it is tagged as “True”, treated as related with the organization.

In the following section, we give the details on how to represent and extract feature, how to build a Maximum entropy classifier and how to utilize Label Propagation algorithm to mine the test data to improve the accuracy of the performance.

4 Supervised Classifier

With the training data, we aim to train a classifier with generic features. The features should not be too general with the preference to tag tweets as True, or too narrow with the preference to tag tweets as False. Furthermore, the features should be generated automatically, with no manual labelling.

4.1 Features Extraction

The features extraction includes tweets features and organization features. In this step, we pay more attention to the organization information, which is expanded with external resources. We induce related homepage, related Wikipedia page, and GoogleSet in the following way to get the features to represent the organization.

Homepage

The URL of each organization is provided in the input. The words in homepage are more related and indicative to the organization. Therefore, these words (removed the stopwords) are chosen to represent the organization. However, some organization webpages are created by java-scripts or even flash, no text information could be extracted from them at present. Therefore, we try to get other external resources.

Wikipedia page

Aim to get higher quality information of the organization, Wikipedia disambiguation page[4] is used. For each organization name, we get some entity candidates related with it through the disambiguation page. If the wiki-webpage of the entity candidate contains the organization's homepage URL, this webpage is treated as a description of the organization, words in this webpage are extracted as features.

Metadata

Meta tags in HTML page provide high quality keywords to represent its webpage. If the webpage has metadata, they are good features to represent organization. However, only a fraction of webpages have this information available.

GoogleSet

GoogleSet provides similar words with the query words. We utilize it to enrich organization information with related words. For example, given words “Yale” and “University”, associated words “Stanford”, “Columbia” are returned. This kind of information is useful, it gives latent semantic category information at some extent.

Capital words

Capital words are more likely to be important words or named entity, we reinforce these words by selecting them as one type of features.

URL

URL in homepage or wiki- webpage is also a strong indictor. If the tweet contains the same URL with homepage or wiki-webpage, it is more possible to be related to the organization.

Corresponding to these types of features for organization, we extract unigrams, bigrams words, capital words and URL from tweets as features.

4.2 Representation

The representation of tweets corresponding to given organization is shown in the following:

$$Vector(T_i, O_k) = \{F_1, F_2, ..., F_n\} \quad (1)$$

Here, T_i is the tweet, O_K is the organization, F_i is one type of features described in the above section. For each F_i, the value is computed as follows.

$$Value(F_i) = \sum_m Wt_m \quad (2)$$

Wt_m is weight of feature t_m, computed by tf*idf or just given {0,1}value. t_m is the co-occurrence feature between F_i and tweet $T_{i.}$. This is similar with the work of (Yerva *et al.*, 2010).

4.3 Maximum Entropy Classifier

The classifier is to classify tweets as True or False with the given feature vector. We aim to train a Maximum Entropy Classifier for this task. The principle of Maximum Entropy Model (Jaynes, 1957) is that the model should maximizes entropy, or "uncertainty" with satisfying all the constraints. This is a straightforward idea that just model what is known, and just keep uniform what is unknown. Here, we utilize all features describe above in this classification task. NLTK[5] tool is used to implement Maximum Entropy Classifier.

5 Semi-Supervised Classifier

The set of organization names in the training and test data are different, this leads to supervised classifier trained on training data is not very effective to the test data. In order to utilize or mine

[4] http://en.wikipedia.org/wiki/xxx_(disambiguation)

[5] http://www.nltk.org/

specific information for certain organization in test data, we adopt one of the classic semi-supervised methods Label Propagation to modify the classification results gotten by Maximum Entropy classifier. We aim to mine the relation among tweets related to one organization in test dataset.

The procedure of Label Propagation Algorithm (Zhu and Ghahramani, 2002) is propagating labels from the labeled vertices (served as seeds) to all the unlabeled ones through the weighed edges in a graph. Larger edge weight will make propagate easier, which means that if the similarity of two nodes is high, they tend to have the same label.

Formally, label distributions are spread across a graph $G=\{V,E,W\}$, where V is the set of n nodes, E is a set of m edges and W is an $n*n$ matrix of weights with W_{ij} as the weight of edge (i, j).

For Label Propagation Algorithm, the seeds selection and graph construction are important. Maximum Entropy classifier gives a confidence value for each tweet which is classified to True or False. Therefore, we choose N True and False tweets tagged by Maximum Entropy classifier as seeds according to its confidence value. In order to get high accuracy of seeds selection, we set N=10.

Each tweet is treated as node, the edge is constructed if two tweets have co-occurrence words, its weight is computed by Cosine similarity.

Our aim is to combine supervised and semi-supervised method to solve this task. Therefore, we select some organizations tweet set to be classified by semi-supervised classification, not all organizations. The selection process is based on the ratio of False tweets in the given organization test set tagged by Maximum Entropy classifier.

$$Ratio(O_k) = Num(False) / Num(O_k) \qquad (3)$$

Here, O_K is the organization, *Num(False)* is number of tweets tagged as False by Maximum Entropy classifier, $Num(O_K)$ is the number of tweets for the given organization. If $Ration(O_K)$ is less than the threshold, the semi-supervised classifier is applied to classify the tweets once more for organization O_K. The other organizations tweets set do not need to be classified by LP.

Here, we make use of JUNTO Label Propagation toolkit[6]. LP_ZGL (Zhu and Ghahramani, 2002), the first label propagation algorithm provided by the toolkit, is chosen in our experiment. All the parameters are set as default.

6 Experiments

6.1 Corpus and Evaluation Metric

We have conducted experiments on the WePS-3 task 2 data. The training data contain about 50 organizations. For each organization, about 400 tweets are provided with tagger {True, False}, corresponding to it is related to the organization or not. The test data also contain about 50 organizations, which are different from those in training data. There is no intersection between training data and test data. For each organization in test data, there are about 400 tweets, which are needed to be classified.

The task is to classify the tweets related or non-related to 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

The performance of classification is shown in Table.1. Here, P+ means the precision of tweets tagged as True, P- means the precision of tweets tagged as False. The Acc means the ratio of tweets tagged with correct tagger.

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

Table 1: Performance of Classification

System	P+	R+	F+	P-	R-	F-	Acc
Baseline 1	1	0	0	0.57	1	0.66	0.57
Baseline 2	0.43	1	0.53	1	0	0	0.43
ME	0.64	0.45	0.44	0.63	0.81	0.66	0.71
ME+LP	0.62	0.54	0.48	0.64	0.65	0.57	0.74
ME+LP+Rule	0.63	0.57	0.49	0.65	0.64	0.58	0.75

Two Baseline systems have been induced, which tags all tweets as related (True) or non related (False). Compare our three systems, the performance is improved by adding LP algorithm and rule step by step. For the F-measure of related tweets, it improves 4%. The recall is improved 9% with a little loss of precision. The whole accuracy is also improved by 3%. It proves that combining supervised and semi-supervised methods is effective. The rule described in Section 3.2: if organization name consisting of more than one word, then the tweet contains full name is treated as related to this organization. Though only one piece of rule is used, it really improves almost all measurement value. It shows that important keywords are good features to distinguish entity.

Compared our systems with WePS participant system, our systems performance are less than the system of (Yerva *et al.*, 2010). Its accuracy value is 0.83. Its system induces manually constructed UserFeedback profile. With only homepage as features, its F-measure of related tweets is 0.3. Different from theirs, our systems are all automatically.

Figure 1 shows the influence of threshold selection for the accuracy, the threshold is described in formula (3).

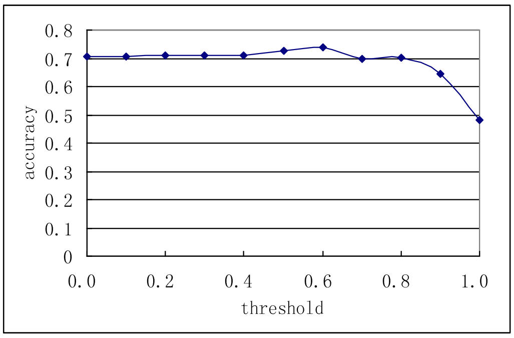

Figure 1: Influence of threshold selection for accuracy

From Figure 1, it is concluded that the performance is decreased by choosing all organizations to be classified again by LP algorithm. As we known, the performance of supervised methods is better than semi-supervised methods. The selection of organization to be classified by LP is important.

7 Conclusion

In this paper, we probe into the problem of finding related tweets to a given organization. This is a challenging task due to the potential organization name ambiguity. This task is more challenging caused by two problems: the tweets and organization contain little information, and the organizations in training data are different with those in test data. We induce external resources to enrich the information of organization. Supervised (ME) and semi-supervised (LP) methods are adopted in two stages to classify the tweets. This is a try to utilize both training and test data for this specific task. Our experimental results on WePS-3 are primary and encouraging, they prove the proposed techniques are effective in performing the task.

There is still a gap needed to be filled by further improving these techniques. For example, the performance can be improved through the semantic expansion of words using Ontology.

References

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

Yerva, S. R., Z. Miklós, and K. Aberer. 2010. It was Easy, when Apples and Blackberries were only Fruits. *Proceedings of the Third Web People Search Evaluation Workshop.*

Tsagkias, M., and K. Balog. 2010. The University of Amsterdam at WePS3. *Proceedings of the Third Web People Search Evaluation Workshop.*

Kalmar, P.. 2010. Bootstrapping Websites for Classification of Organization Names on Twitter. *Proceedings of the Third Web People Search Evaluation Workshop.*

Yoshida, M., S. Matsushima, S. Ono, I. Sato, and H. Nakagaw. 2010. ITC-UT: Tweet Categorization by Query Categorization for On-line Reputation Management. *Proceedings of the Third Web People Search Evaluation Workshop.*

Zhu, X. J., and Z. Ghahramani. 2002. Learning from Labeled and Unlabeled Data with Label Propagation. *Technical Report CMU-CALD-02-107, Carnegie Mellon University.*

Amigó, E., J. Artiles, J. Gonzalo, D. Spina, B. Liu, and A. Corujo. 2010. WePS-3 Evaluation Campaign: Overview of the Online Reputation Management Task. *Proceedings of the Third Web People Search Evaluation Workshop.*

Jaynes, T.. 1957. Information Theory and Statistical Mechanics. *Physics Reviews*, (106), 620-630.

Perez-Tellez, F., D. Pinto, J. Cardiff, and P. Rosso. 2011. On the Difficulty of Clustering Microblog Texts for Online Reputation Management. *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT*, pp. 146-152.

Dan, O., J. Feng, and B. D. Davison. 2011. A Bootstrapping Approach to Identifying Relevant Tweets for Social TV. *Proceeding of fifth International AAAI Conferenc Weblogs and Social Media (ICWSM),* pp. 462-465.