An Adaptive Method for Organization Name Disambiguation with Feature Reinforcing

Shu Zhang[1], **Jianwei Wu**[2], **Dequan Zheng**[2], **Yao Meng**[1] **and Hao Yu**[1]
[1] Fujitsu Research and Development Center
Dong Si Huan Zhong Rd, Chaoyang District, Beijing and 0086, China
{zhangshu, mengyao, yu}@cn.fujitsu.com
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 an online social networking, which has become an important source of information for marketing strategies and online reputation management. 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 both from organization and the tweets. We mine organization information from web sources to train a general classifier. Further, we mine tweets information. We train an adaptive classifier for a given organization name with more features derived from twitter messages labeled by the general classifier. The experiments on WePS-3 show mining web sources to enrich organization are effective. The adaptive classifier trained for a given organization is promising.

1 Introduction

Twitter is an online social networking and microblogging service, which rapidly gained worldwide popularity, with 140 million active users as of 2012[1], generating over 340 million tweets and handling over 1.6 billion search queries per day[2]. People share their opinions on almost anything on Twitter, such as news, governmental policies, products and companies. Therefore, Twitter becomes an important information resource for the purpose of marketing strategies and online reputation management. 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, such as product, company, or certain event. This work 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*" could be used to refer river or company. Therefore, when the entity name is ambiguous, filtering spurious name matches is important to accurate detection and analysis of contents that people say about the given entity.

This paper focuses on finding related tweets to a given organization. Assuming that tweets are retrieved by the query of organization name, such as "*apple*", the task is to identify whether a tweet is relevant to the target organization ("*Apple Inc.*") or not. Yerva *et al.* (2010) adopt support vector machines (SVM) classifier to classify tweets with external resources. Yoshida *et al.* (2010) classify

[1] http://blog.twitter.com/2012/03/twitter-turns-six.html

[2] http://engineering.twitter.com/2011/05/engineering-behind-twitters-new-search.html

26 th Pacific Asia Conference on Language, Information and Computation pages 237-245

organization names into "organization-like names" or "general-word-like names" categories, classify tweets by rules. Kalmar (2010) adopts bootstrapping method to classify the tweets.

This task is challenging owing to the fact of lacking sufficient information. A tweet contains less than 140 characters and is often freely written. Therefore the tweet is short and informal. It does not provide sufficient word occurrence or context shared information for effective similarity measure (Phan *et al.*, 2008). Furthermore, the representation of each organization is also an obstacle. Different from conventional word disambiguation, there is no authoritative source which lists all possible interpretations of an organization name. The information gotten from the homepage of organization is limited. It is difficult to cover the word occurring in tweets which are related to the given organization.

Aim to process any organization names but not one or some given organization names, the organization names in training data are different from those in test data. This leads that we could not train a classifier to a certain organization. It also makes the task more difficult than conventional classifying task.

In this paper, we propose an adaptive method for organization name disambiguation. We build a general classifier with the training data. Then we use the general classifier to label unlabeled twitter messages of a given organization. With more features derived from these twitter messages, we train an adaptive classifier to a given organization. The major contributions of our approach are as follows:

- Try to mine organization information from web sources, such as Wikipedia, linked pages and related pages. This is a way to solve the problem of insufficient information.
- Train an adaptive classifier for a given organization name with more features derived from twitter messages labeled by general classifier. This is a way to let the classifier more suitable for a 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 presents supervised methods to classify tweets based on information from web sources. Section 5 introduces adaptive method to classify the tweets based on derived features. Section 6 gives the experiments and results. Finally section 7 summarizes this paper.

2 Related Work

Online social networks such as Twitter have attracted much interest from the research community. With little information contained in each tweets, it is a challenge for monitoring and analyzing them. There are some relevant works studied recent years.

Meij *et al.* (2012) add semantics to tweets by automatically mapping tweets to Wikipedia articles to facilitate social media mining on a semantic level. Liu *et al.* (2011) focus on NER on tweets and use a semi-supervised learning framework to identify four types of entities. Sriram *et al.* (2011) focus on classifying twitter messages to a predefined set of generic classes such as News, Events, Opinions, Deals, and Private Messages.

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).

In WePS-3, 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, including 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. 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.

General classifier of our work is similar to the research of (Yerva *et al*., 2010) in the manner of constructing profiles for each organization and

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

forming general features. Different from theirs, we try to introduce different kinds of web pages to fully represent the organization as far as possible.

3 Overview

3.1 Problem Statement

Given a set of tweets and an organization name, the goal is to decide if each tweet in the set talks about this organization.

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. Table 1 shows the examples of tweet disambiguation for the company "Cadillac".

	Tweet content	Tag
1	On Sale: 2004 Hotwheels Crank Itz 3/5 Cadillac Escalade	TRUE
2	Update: Cadillac CTS-V vs BMW M5 Performance Testing.....	TRUE
3	#nowwatching cadillac records while I'm finishing my paper	FALSE
4	.....founded in 1701 by the Frenchman Antoine de la Mothe Cadillac	FALSE

Table 1: Examples of tweet ambiguity for the company name "Cadillac"

3.2 Our Method

Overcome the challenges of this task, we import web resources to enrich more information about the organization, such as homepage, Wikipedia page, related webpage, and unrelated webpage. With the general features extracted from these resources and training data, we train a general classifier.

Given an organization name in test data, we label the tweets by general classifier first. More features are derived from these tweets. The adaptive classifier for a given organization is trained with both the general features and derived features. Figure 1 gives an overview of our method.

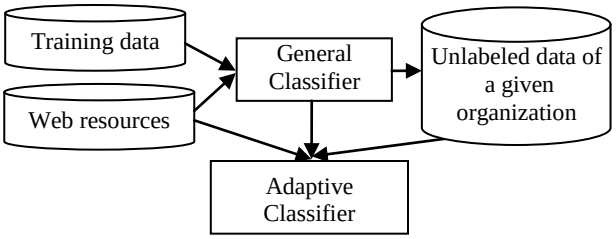


Figure 1. Overview of our method

4 General Classifier

From the input information, we may get the information related to the given organization from homepage URL. The information from homepage is important. However its coverage is limited. The tweet and organization homepage alone contain very little sharing information for effective similarity measure. Therefore, we try to mine web sources to enlarge the coverage of information related to the organization.

There is another problem. In this task, we have a training set corresponding to a few organization names. However, the organization names in test set do not appear in training set. This scenario can be seen as in-between supervised and unsupervised learning. The conventional lexical level features are not effective for classifying different organization names, because these organizations may belong to different domains. Therefore, we try to generate more general features from the web sources, train a classifier on training data, and classify the tweets corresponding to the unseen organization names in test set. We adopt Maximum Entropy, Support Vector Machine, and Naive Bayes methods to train the classifier.

4.1 Mine Organization Information from Web Sources

Here, we aim to mine the following web sources to get the information about the given organization.

Homepage

It is natural to regard that the organization's web site is indicative to represent the organization. We crawl through web pages from the homepage in maximum depth of 2.

However, some homepages are edited by javascripts or even flash, from which no valuable

text could be extracted. At present, we discard these homepages.

Wikipedia related webpage

As a well organized and freely available knowledge, Wikipedia provide high quality information for some entity. Because lexical ambiguity exists, we utilize Wikipedia disambiguation page [4], which provides some candidates for a given entity name. If the wiki-webpage of an entity candidate contains the organization's homepage URL, we believe that this webpage is related to the organization. However, we can't find the related wiki-webpage for all of the organizations, because of the limited coverage of wikipedia or homepage URL mismatch.

Wikipedia unrelated webpage

Once finding Wikipedia related page, the remaining candidates of the disambiguation page are selected as Wikipedia unrelated pages. These web pages may contain the information that indicates the other meaning of organization name.

Figure 2 shows an example of Wikipedia disambiguation page of "http://en.wikipedia.org/wiki/Apple_(disambiguation) ". In this webpape, Apple Inc is the company we cared as Wikipedia related webpage, the others are treated as unrelated webpages.


Companies

- Apple Inc., a consumer electronics and software company founded in the 1970s
- Apple Corps, a multimedia corporation founded in the 1960s by The Beatles
- Apple Bank, an American bank in the New York City area

Films

- *The Apple* (1980 film), a 1980 musical science fiction film
- *The Apple* (1998 film), by Samira Makhmalbaf
- *Apples* (2012 film), a 2012 adaption of Richard Milward's novel

Television

- "The Apple" (Star Trek: The Original Series), a 1967 second season episode

Music

- *Apple* (album), a 1990 album by Mother Love Bone
- Apple (band), a British psychedelic rock band
- The Apples, an early 90s Scottish indie-dance band
- The Apples (Israeli), a mid 2000s Israeli funk band
- Apple Records, record label founded by The Beatles

Figure 2. An example of Wikipedia disambiguation webpage

Related webpage

Google provides the search key word "related", which is used to find related or similar web page for a given URL. For example, input a query "related: http://www.apple.com", Google would return many web sites of other electronic companies, such as HP, DELL, and SONY as shown in Figure 3. These web pages contain the category information related to the given organization, which enlarge the coverage of organization information in some extent. Here, we collect top-100 retrieval result as related web pages.


Dell Official Site - The Power To Do More | Dell
www.dell.com/
2010年9月27日
Dell provides technology solutions, services and support. Visit Dell.com for Laptops, Netbooks, Tablet PCs ...

HP - United States | Laptop Computers, Desktops, Printers, Servers ...
www.hp.com/ - 翻译此页
Research or buy HP printers, desktops, laptops, servers, storage, enterprise solutions and more at the Official Hewlett-Packard Website.

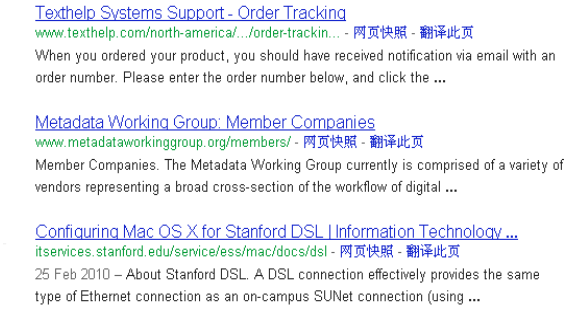
AppleInsider | Apple news and rumors since 1997
www.appleinsider.com/ - 网页快照 - 翻译此页
Home · Reviews · Forums · Price Guides · Current-Gen Macs · Previous-Gen Macs · Apple Wireless Devices · iPads, iPhones & iPods · Used Apple Products ...

Microsoft Corporation: Software, Smartphones, Online, Games ...
www.microsoft.com/ - 网页快照 - 翻译此页
Get product information, support, news from Microsoft.

Figure 3. An example of related webpage

Link webpage

Similar with related web pages, Google provides another search key word "link", which is used to find web pages linked to a specified URL. For example, input a query "link: http://www.apple.com", we access to a wider variety of results which contain a URL of "http://www.apple.com", as shown in Figure 4. We think the web pages linked to given URL are information extension of organization, may have some relationship with the organization. Top-100 retrieval results are collected as link web pages.


Texthelp Systems Support - Order Tracking
www.texthelp.com/north-america/.../order-trackin... - 网页快照 - 翻译此页
When you ordered your product, you should have received notification via email with an order number. Please enter the order number below, and click the ...

Metadata Working Group: Member Companies
www.metadataworkinggroup.org/members/ - 网页快照 - 翻译此页
Member Companies. The Metadata Working Group currently is comprised of a variety of vendors representing a broad cross-section of the workflow of digital ...

Configuring Mac OS X for Stanford DSL | Information Technology ...
itservices.stanford.edu/service/ess/mac/docs/dsl - 网页快照 - 翻译此页
25 Feb 2010 – About Stanford DSL. A DSL connection effectively provides the same type of Ethernet connection as an on-campus SUNet connection (using ...

Figure 4. An example of link webpage

4.2 General Features and Representation

Once we have collected the above five kinds of web pages, the crawled web pages are preprocessed, including removing HTML tags, filtering stop words, and stemming. Finally, all unigrams and bigrams are chose to represent the

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

organization. We extract the following four types of information to construct profiles, in fact each profile can be treated as a set of key words.

Unigram profile: $P_u = set\{uigram\}$

Bigram profile: $P_b = set\{bigram\}$

Metadata profile: $P_m = set\{word\}$

URL profile: $P_{url} = set\{host\_name\}$

We construct 22 binary general features as follows.

$$F(T_i, Org) = \{\underbrace{F_u^h, F_b^h, F_m^h, F_{url}^h}_{\text{home\_page}}, \underbrace{F_u^w, ..., F_{url}^w}_{wiki\_page}, \underbrace{F_u^{nw}, ..., F_{url}^{nw}}_{neg\_wiki\_page}, \underbrace{F_u^l, ..., F_{url}^l}_{link\_page}, \underbrace{F_u^r, ..., F_{url}^r}_{related\_page}, \underbrace{H_1, H_2}_{heuristics}\}$$

$$T_i = set\{key\}$$

$$F_j = \begin{cases} 1, if\ T_i \cap P_j \neq NULL \\ 0, else \end{cases}$$

Where T_i represents the *i*-th tweet, *Org* is the given organization, and P_j is a profile. F_j is the weight of the corresponding feature. T_i use the unigram, bigram and URL as the key to represent tweet corresponding to different profiles of organization.

For different organization names, the given organizations are needed to have their own profiles from the five given web sources. We use the similarity between the tweet and organization profiles as the general features. These features are stable for different organizations. However, the classifier built with conventional lexical features is highly dependent on organizations, because it has different weights of lexical features for different organizations. In this task, the set of organization names in training and test data set are different. Therefore, general features are more suitable than lexical features for building a classifier with training data.

From these general features, we measure the similarities between a tweet and a given organization on a level of different web sources, but not lexical level.

In addition, we also utilize the following two heuristic rules:

H1: if an organization name have multiple words, we set value as 1, else set as 0;

H2: if a tweet contains the full organization name, we set value as 1, else set as 0;

We think organization name with multiple words may contain more information. For example, "Yale University" contains more semantic information to distinguish it from other entity.

So far, we have formed general features, which are not organization specific. Each tweet is represented by this kind of features would have the same distribution between training and test set. So, traditional supervised classifiers could be applied and have good generalization performance on unseen data.

4.3 Supervised Classifiers

Here, we train three classical supervised classifiers with the general features gotten from the web sources, with the aim to get general classifiers to classify the tweets.

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 is that the model should maximize 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 described above in this classification task. NLTK [5] tool is used to implement Maximum Entropy Classifier.

Support Vector Machine

Support Vector Machine (SVM) is a popular machine learning approach. Based on the structural risk minimization of statistical learning theory, SVM finds an maximum-margin hyperplane to separate the training examples into two classes. Due to maximum-margin preventing over-fitting in high-dimensional data, SVM usually achieves good performance on a range of tasks.

We use SVMLight [6] toolkit to achieve the classification result. RBF kernel function is used and all the other parameters are set to their default values.

Naive Bayes Classifier

The Naive Bayes Classifieris based on Bayesian theorem. Though it is simplicial, Naive Bayes Classifier has been proved very effective for text

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

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

categorization. We use the Naive Bayes Classifier provided by the NLTK toolkit.

5 Adaptive Classifier

In this task, organization names in training data are different from those in test data. In Section 4, we train supervised classifier with general features on training data. In this section, we aim to get an adaptive classifier to a certain organization in test data. The adaptive classifier is trained with more features gotten from the tweets in the test set for a given organization.

5.1 Adaptive Process

The adaptive process includes three parts: (1) get labeled data, (2) derive more features, and (3) train classifier. The detail is given in the following algorithm.

Algorithm: Adaptive process

Input: general classifier(GC) and Tweet set(TS) of a given organization

Output: adaptive classifier

Algorithm:

(1) Label TS using GC, and get result(GR) ;

(2) Derive features from GR, choose feature type and extract feature using feature selection method ;

(3) Train adaptive classifier (AC) to a certain organization, using both general features and derived features with GR.

Here, we try two ways to get the tweet set of a given organization. One is to use the data in test set directly, the other is to crawl tweets from twitter with organization name as query. To different organization name, the scale of the retrieved tweets from twitter is more than 2,000, which is larger than the test data with about 400 tweets for a given organization name.

We use general classifier to label the tweets of a given organization. From the results, we could derive more features and train an adaptive classifier.

For training the adaptive classifier, we use both general features and derived features, with the aim of utilizing both the information from web sources and data set of a given organization.

5.2 Derived Features

Lexical level features are important for classification task. We do not use lexical features for general classifier because they are changing for different organizations. The weights of lexical features are quite different for different organizations. However when the organization is given, lexical features could distinguish related or unrelated tweets effectively.

Feature type

We adopt two types of features: one is the unigram word unit, the other is 4-gram character unit.

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.

Feature selection

The features derived from the labeled tweets are large scale and contain much noise. We need adopt feature selection method to get more effective features.

Here, we first select the features which have more than five times occurrences in tweet set of a given organization. Then we adopt Information Gain (IG) method to select top N features with high value of IG. IG is one of the classical feature selection methods. We set N as 2,000.

6 Experiments and Results

6.1 Corpus and Evaluation Metric

We have conducted experiments on the WePS-3 task 2 data. The training data contain about 50 organizations with about 400 tweets for each organization. The test data also contain about 50 organizations. There is no intersection between training and test data.

The task is to classify the tweets related or non-related to the given organization, it belongs to classification task. In details, there are four categories for the tweets in evaluation phase: true positive(TP), false positive(FP), true negative(TN), false negative(FN). Therefore, we measure the performance by accuracy, precision, recall and F-measure.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision^{+} = \frac{TP}{TP + FP} \qquad \text{Recall}^{+} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$Precision^{-} = \frac{TN}{TN + FN} \qquad \text{Recall}^{-} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$F - Measure^{+} = \frac{2 * Precision^{+} * \text{Re}call^{+}}{Precision^{+} + \text{Re}call^{+}}$$

$$F - Measure^{-} = \frac{2 * Precision^{-} * \text{Re}call^{-}}{Precision^{-} + \text{Re}call^{-}}$$

6.2 Results and Analysis

We testify our proposed methods from the following aspects:

- The effectiveness of general classifier built with training data and information from web sources
- The influence of information gotten from different web sources for the performance of general classifier
- The effectiveness of adaptive classifier with derived features and unlabeled tweets of a given organization

Performance of general classifier

First, we testify the performance of supervised classifiers built with training data and information from web sources.

Table 2 shows their performance and also lists the performance of the state of art methods. Top_1, Top_2 and Top_3 are the 3 best system results in Weps-3 task 2 evaluation. $BASELINE_R$, $BASELINE_{NR}$ are the baselines with arbitrary prediction that tag all tweets just related or non-related respectively.

	ACC	F +	F -
NB	**0.7508**	**0.5823**	0.6444
ME	**0.7510**	0.5375	0.6755
SVM	0.7383	0.5153	0.6506
Top_1	**0.8267**	0.6264	0.5606
Top_2	0.7491	0.4935	0.5651
Top_3	0.7312	0.5062	0.4683
$BASELINE_{NR}$	0.5652	0.0000	0.6563
$BASELINE_R$	0.4348	0.5274	0.0000

Table 2: Performance of supervised methods and other methods

In Table 2, the accuracy of $BASELINE_{NR}$ is higher than that of $BASELINE_R$, which shows that there are more unrelated tweets in the whole test data. The performances of Naive Bayes (NB), Maximum Entropy (ME) and Support Vector Machines (SVM) have similar values of accuracy. They are much higher than those of $BASELINE_{NR}$ and $BASELINE_R$. It proves that adopting some methods to disambiguate tweets is necessary.

Our proposed methods have the similar accuracy values with Top_2 and Top_3. It proves that proposed supervised classifiers, built with training data and information from web sources, are effective for this task.

The accuracy value of our methods is lower than that of Top_1. Its accuracy value is nearly 0.83, Top_1 method adopts manually constructed user feedback profile. With only homepage as features, its accuracy is about 0.66, which is similar with performance of our methods shown in Figure 5. Different from theirs, our methods are all automatically.

Compare with ME and SVM classifiers, NB classifier has better performance in F+ values. F+ value is important to measure the ability of finding the related tweets to a given organization.

Influence of different web sources for performance of general classifier

We select NB classifier to find the influence of information gotten from different web sources. Figure 5 lists the performance of NB classifier built with information gotten from only one of five different web sources.

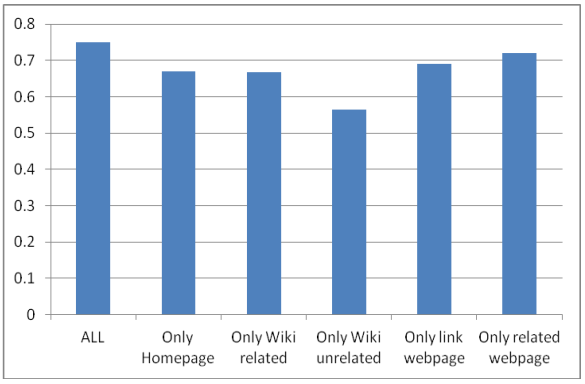


Figure 5. Accuracy of supervised methods (NB classifier) with different web sources

From Figure 5, we can see that the accuracy of classifier combining these five web sources is

highest, which means the combination of five web sources is effective and feasible. This also shows that mining web sources is an effective way to enhance the performance of disambiguation.

Among the five classifiers built with features gotten from only one of web source, the accuracy of classifier built with information from related webpage is much higher than that of others. That means the related webpage containing the category information related to the given organization is much useful, which enlarge the coverage of organization information in some extent.

The accuracy of classifier built with only link webpage is also higher than that of homepage or wiki unrelated webpage. This shows link webpage and related webpage give more information about the given organization, our proposed web sources is effective for this task.

However, the performance of classifier built with the features gotten only from homepage is not as good as expectation. This may be caused by the information limitation, which could not cover the information of tweets. The focus of tweets may be different from that of homepage.

The accuracy of classifier built with information from Wiki unrelated webpage is the lowest. Our purpose of importing Wiki unrelated webpage is to mine the negative information about a given organization. Therefore, it should not be used by only itself. It is better to combine wiki unrelated webpage with other web sources.

Performance of adaptive classifier

We select NB classifier as the general classifier to label the tweets of a given organization name. Then we utilize them to train an adaptive classifier for this given organization. As described in Section 5.1, we adopt two ways to get the tweets of a given organization. One is to use test data, which is tagged as Adaptive-T. The other is to retrieve tweets from Twitter, which is tagged as Adaptive-U. The scale of unlabeled data is shown in Table 3. The performances are shown in Table 4.

	Number of tweets
Tweets of test data	~400
Tweets from Twitter	2,500-8,000

Table 3: Number of unlabeled tweets of one given organization

	ACC	**F+**	**F-**
NB	0.7508	0.5823	0.6444
Adaptive-T	0.7629	0.5676	0.6334
Adaptive-U	**0.7697**	0.5982	0.6618

Table 4: Performance of adaptive classifier

Table 3 shows that the scale of tweets from Twitter is much larger than that of test data. The size of tweets from Twitter is ranged from 2,500 to 8,000. This is dependent on whether the organization is hot point or not.

From Table 4, we can see that the accuracies of both adaptive classifiers are higher than that of NB classifier, which show that the proposed adaptive process is effective. With unlabeled data, derived more lexical features in adaptive process is one way to improve the performance of disambiguation.

The scale of tweets retrieved from Twitter is much larger than that of test data. Therefore, the coverage of lexical features of adaptive-U is larger than that of adaptive-T, the performance of adaptive-U is better than that of adaptive-T.

Besides accuracy, F+ and F- of adaptive-U are also higher those of NB classifier. This shows that mining large scale of unlabeled tweets is an effective way to get more information about a given organization.

7 Conclusion

In this paper, we probe the problem of organization name disambiguation on twitter information. We propose an adaptive method for organization name disambiguation. We build a general classifier with the training data and different web sources. Then we use the general classifier to label unlabeled twitter messages of a given organization. With more features derived from these messages, we train an adaptive classifier to a given organization. The experiments on WePS-3 show that the general classifier is effective for this task. The adaptive classifier improves the performance of general classifier, especially with a large scale of tweets gotten from Twitter.

In the future, we will try to select more features in the adaptive process, and find their influences for the performance of adaptive classifier. Furthermore, we will try to propose some methods to reduce the noise from both tweets and organization information.

References

Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, Murat Demirbas. 2011. Short Text Classification in Twitter to Improve Information Filtering. Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 841-842.

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

Edgar Meij, Wouter Weerkamp, Maarten D. Rijke. 2012. Adding Semantic to Microblog Posts. Proceedings of the 4th ACM Web Search and Data Mining, 563-572.

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

Fernando Perez-Tellez, David Pinto, John Cardiff, Paolo 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, 146-152.

Haewoon Kwak, ChanghyunLee, Hosung Park, Sue Moon.2010. What is Twitter, a Social Network or a news Media? Proceedings of the 19th International Conference on World Wide Web, 591-600.

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

Miguel García-Cumbreras, Manuel García-Vega, Fernando Martínez-Santiago, José M. Perea-Ortega. 2010. SINAI at WePS-3: Online Reputation Management. Proceedings of the 3rd Web People Search Evaluation Workshop

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

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

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

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

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

Xuan-Hieu Phan, Le-Minh Nguyen, Susumu Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. Proceedings of the 17th International Conference on World Wide Web, 91-100.