ITC-UT: Tweet Categorization by Query Categorization for On-line Reputation Management

Minoru Yoshida, Shin Matsushima, Shingo Ono, Issei Sato, and Hiroshi Nakagawa

University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033
{mino,masin,ono,sato,nakagawa}@r.dl.itc.u-tokyo.ac.jp

Abstract. This paper describes our system, called ITC-UT, for the task-2 (on-line reputation management task) in WePS-3. Our idea is to categorize each query into 3 or 4 classes according to how much the tweets retrieved by the query contain the "true" entity names that refer to the target entity, and then categorize each tweet by the rules defined for each class of queries. We show the evaluation results for our system along with the details of results of query categorization.

Keywords: Organization Name Disambiguation, Two-Stage Algorithm, Naive Bayes, Twitter

1 Introduction

This paper reports the algorithms and results of the ITC-UT (Information Technology Center, the University of Tokyo) team for the WePS-3 task-2 (on-line reputation management task.) The supposed situation of this task is where you search reputation of some organization in Twitter. Assuming that tweets are retrieved by the organization name query, the problem is to decide whether each organization name found in each tweet represents the target organization or not (such as "Apple PC" for the former and "Apple Pie" for the latter for the query "Apple".) This is one type of name disambiguation problems that have been extensively studied through previous WePS workshops[1, 2]. However, the current task setting is challenging because generally each tweet is small and provides little context for disambiguation.

Our algorithm to solve this problem is based on the intuition that organization names can be classified into "organization-like names" and "general-word-like names", such as "McDonald's" for the former and "Pioneer" for the latter. This intuition is supported by the fact that the ratio of TRUE[1] (or FALSE) tweets in the training data vary widely from entity to entity. For example, over

[1] TRUE indicates that the tweet mentions the target organization (as defined in the next section). FALSE indicates the opposite.

98% of tweets were labeled TRUE for entity "nikon", while the ratio for entity "renaissance technologies" (for which the query term was "Renaissance") was under 1%. Our strategy is to make aggressive use of such unbalance by predicting whether each query in the test set is biased towards TRUE or FALSE as described in detail in section 3.1. Then the heuristic rules suited for the bias of query are applied to categorize the tweets. For instance, if a query is highly likely to be an organization name, each tweet is labeled TRUE unless some strong evidences indicate the opposite. The detail is described in section 3.2.

2 Task Definitions

In this section, we briefly give the definition of the task required for the description of our algorithm. Both the training and test data contain the entity name (e.g., "marriott international"), the query term used to retrieve tweets (e.g., "Marriott"), the URL of the home page for the entity, and 700 tweets (per entity name) retrieved by the query term. The training data also contain the label "TRUE" or "FALSE" for each tweet that indicate whether the tweet mentioned the entity or not. The task is to predict whether each tweet in the test data (provided with no label) are TRUE (i.e., mentions the entity) or FALSE (i.e., doesn't mention the entity.)

3 Algorithm

As mentioned above, our algorithm is mainly divided into two stages: the query categorization stage (stage 1) and the tweet categorization stage (stage 2). In this section, we describe each stage in more detail.

3.1 Stage 1: Query Categorization

The first stage categorizes each query into 3 or 4 classes according to the confidence of "how the query indicates the given organization if no contexts are given".

For training data, the class of each query was determined by the ratio of the number of TRUE tweets (represented by t) to the number of all tweets for the query. We used two different configurations for the number of classes: 3 and 4. In the 3-class settings, each query is categorized into:

class 1: TRUE-biased queries: if $t > \theta_1$,
class 2: FALSE-biased queries: if $t < \theta_2$,
class 3: neutral queries: otherwise.

On the other hand, in the 4-class settings, each query is categorized into:

class 1: TRUE-biased queries: if $t > \theta'_1$,
class 2: FALSE-biased queries: if $t < \theta'_2$,
class 3: neutral queries: if $\theta'_3 < t \leq \theta'_1$,

class 4: weakly FALSE-biased queries: otherwise.

The threshold values θ_i and θ'_i were determined manually by looking at the training data. The values were $\theta_1 = 0.66..$ and $\theta_2 = 0.33..$ for 3-class labeling, and $\theta'_1 = 0.9$, $\theta'_2 = 0.1$, and $\theta'_3 = 0.5$ for 4-class labeling.

For categorization, we did not use linguistic features (e.g., frequent words in tweets) other than very simple ones by pattern matching (such as "Is an acronym?" feature described below) because useful linguistic features for classification seem to be different for different entities and it is difficult to find the features common between training and test data. Instead, we made an extensive use of meta-data such as URLs. The categorization was performed by the simple Naive Bayes classifier (in the Weka[2] toolkit) with following 6 binary features.

Is the query identical to the entity name? This feature value is true for query "Apple" for entity "Apple" and false for query "Amazon" for entity "Amazon.com", for example. This feature is introduced based on the intuition that the difference between the query and the entity name suggests that the entity requires the full name to be specified, such as "Delta Holding" which may tend to be confused with other organizations including "Delta Air Lines" when the query "Delta" is used.

Does the domain name in URL include the query or entity name? This feature value is true if, for example, the URL can be described by the regular expression `http://(www.)?apple.[a-z]/` for the query "Apple". This feature being true may indicate that the organization has an original domain, and therefore a not so minor organization.

Does Wikipedia have "disambiguation page" for the query? This feature is introduced based on the intuition that highly ambiguous names, for which the disambiguation task is difficult, might have a disambiguation page in Wikipedia (www.wikipedia.org).

Is the query an acronym? This feature is based on the observation that acronyms tend to have high ambiguity because they have typically only 2 or 3 characters and therefore many different concepts are expressed by the same acronym.

Does the given URL indicate the top page of Web search results? If the given entity is a major concept represented by the query word, the URL for the entity will come to the first of the search result list if we enter the query to an internet search engine, in which case the feature value is set to "true."

Is the query an entry of a dictionary? This feature is introduced to detect whether the query word is a general word or not. If the former is the case, it will be a risk of the query being used not as the specific organization name, but as some general words.

[2] http://www.cs.waikato.ac.nz/ml/weka/

3.2 Stage 2: Tweet Categorization

Stage 2 categorizes each tweet into "mentioning on the organization" (TRUE) or not (FALSE). The categorization is decided by simple heuristic rules defined for each class of queries.

The system obtains Part of Speech (POS) tags and Named Entity (NE) labels of the queries in each tweet by using Stanford POS tagger[3] and NE Recognizer[4]. Each tweet is categorized by rules that use these extracted POS and NE labels. These rules are defined for each class of queries as follows.

Class 1: TRUE-Biased Queries Each tweet for this class is categorized into TRUE unless it is strongly suggested that, by the following rules, the query represents something other than organizations.

1. If the NE tag of the query is a "PERSON" or "LOCATION", label FALSE.
2. Otherwise, label TRUE.

Class 2: FALSE-Biased Queries On the contrary to the class 1 rules, the tweet for this class of queries is categorized into FALSE unless it is strongly suggested, by the following rules, that the query does represent the orgznization.

1. If the entity name consists of two or more words (such as "Cisco Systems"), and it is contained in the tweet, label TRUE.
2. If the tweet contains the URL for the entity, label TRUE.
3. Otherwise, label FALSE.

Class 3: Neutral Queries Rules for the tweets for the queries of class 3 are the same as the rules for class 1 except that we add another rule (the second one) to detect FALSE tweets because the ratio of FALSE tweets may be larger than the class 1. The rules for class 3 therefore are defined in the following way.

1. If the NE tag of the query is "PERSON" or "LOCATION", label FALSE.
2. If the POS tag of the query is not a proper noun, label FALSE.
3. Otherwise, label TRUE.

We have another version of the rules that replaces the second rule with the following one. This difference of the versions adjusts the filtering power of the additional rule where the above one is stronger (filtering out (i.e., labeling FALSE) more tweets) and the below one is weaker (filtering out less tweets)[5]. We call the original version of rule 2 *the strong filter* and the alternative one *the weak filter*.

2. If the POS tag of the query is not a noun, label FALSE.

[3] http://nlp.stanford.edu/software/tagger.shtml
[4] http://nlp.stanford.edu/software/CRF-NER.shtml
[5] Note that proper nouns are also nouns.

Class 4: Weakly FALSE-Biased Queries This class is optional and the following rules are used. The rules for this class are the same as the rules for class 2 except that we add another rule (the third one) to find more TRUE tweets because more TRUE tweets are expected for this class than class 2.

1. If the entity name consists of two or more words and it is contained in the tweet, label TRUE.
2. If the tweet contains the URL for the entity, label TRUE.
3. If the NE tag of the query is "ORGANIZATION", label TRUE.
4. Otherwise, label FALSE.

System Parameters We used four different configurations for submission, resulting in four runs and outputs. The four configurations are listed below.

ITC-UT_1: used 3 classes and the strong filter (proper noun) for the class 3 rules.
ITC-UT_2: used 3 classes and the weak filter (noun) for the class 3 rules.
ITC-UT_3: used 4 classes and the strong filter (proper noun) for the class 3 rules.
ITC-UT_4: used 4 classes and the weak filter (noun) for the class 3 rules.

4 Experimental Results

We participated in the WePS-3 evaluation campaign with the four systems mentioned above. In this section, we report the performances of our methods. As described above, the systems are different in their rules for tweet categorization and the number of classes for query categorization. These specifications are again shown in Table 1.

The accuracy, precision, recall and F-measure of each method were calculated both for positive and negative examples. We show those values of our algorithms and the top system (indicated by "LSIR,EPF_1") in Table 2.

Among our methods, ITC-UT_1 achieved the best accuracy, which took the second position in the evaluation campaign. When we introduced "weakly FALSE-biased class", the performance degraded in most of the measures while only recall for negative example increased in both cases. It is natural that recall for negative example increased when we introduced "weakly FALSE-biased class" because tweets in this class are more likely to be classified to FALSE than the neutral class. Performance drop in the other measures suggests that the number of queries categorized to "weakly FALSE-biased class" was unnecessarily large, which may be because the conditions to specify "weakly FALSE-biased class" for the training data was too loose.

As shown in the table, when the rule 2 for class 3 changed from the strong filter (proper noun) to the weak filter (noun), most of values degraded while only recall for positive example increased. The "weak filter" contributes to save (i.e., label TRUE) more TRUE tweets (i.e., true positives) while it also saves more

Table 1. Specification of Each Methods

Method	rules	number of category
ITC-UT_1	NE	3
ITC-UT_2	Noun	3
ITC-UT_3	NE	4
ITC-UT_4	Noun	4

Table 2. Performances of Methods

Method	Accuracy	Precision (Positive)	Recall (Positive)	F-measure (Positive)	Precision (Negative)	Recall (Negative)	F-measure (Negative)
LSIR,EPFL_1	**0.83**	0.71	**0.74**	**0.63**	**0.84**	0.52	0.56
ITC-UT_1	0.75	**0.75**	0.54	0.49	0.74	**0.60**	**0.57**
ITC-UT_2	0.73	0.74	0.62	0.51	0.74	0.49	0.47
ITC-UT_3	0.67	0.70	0.47	0.41	0.71	0.65	0.56
ITC-UT_4	0.64	0.69	0.55	0.43	0.70	0.55	0.46

FALSE tweets (i.e., false positives.) The result showed that the increase of the former (true positives) was surpassed by the increase of latter (false positives).

We also compared our methods with the top system in the campaign (LSIR,EPFL_1). Our algorithm tend to show higher precision for positive examples and higher recall for negative examples, which implies our methods are biased to labeling FALSE. We think that our tweet classification rules, especially for class 3 ("neutral class"), leaves much room for improvement.

In Table 3 we show the classification results in the first stage. Roughly speaking, the result indicates that our algorithms could catch the biases of each query. However, it is not fully obvious whether each query was successfully labeled.

Note that labeling of the training queries was different between 3-class and 4-class settings because the threshold values are different between them. We show the detailed results of labeling of training queries in Table 4. Currently, we did not perform any adjustment to tune the threshold values for labeling of the training queries to be better fit to the stage-2 rules for each class of queries. We think these threshold values of labeling of training queries can be improved by, for example, cross validation on the training data or simply maximizing accuracy of training data.

5 Conclusions

This paper reported the ITC-UT system for tweet categorization for the on-line reputation management task, which uses the 2-stage algorithm that categorizes each query in the first stage, and categorizes each tweet in the second stage using the rules customized for each class of queries. Our categorization rules are rather simple, therefore they still leave for improvement. For example, we can adjust the threshold values used in stage-1 to label the queries more appropriately for fitting to the stage-2 rules. We think we can also improve the results by using more sophisticated rules for tweet categorization for each classified class of queries.

References

1. Artiles, J., Gonzalo, J., Sekine, S.: The SemEval-2007 WePS evaluation: Establishing a benchmark for the web people search task. In: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007). pp. 64–69 (2007)
2. Artiles, J., Sekine, S., Gonzalo, J.: Web people search: results of the first evaluation and the plan for the second. In: Proceeding of the 17th international conference on World Wide Web (WWW '08). pp. 1071–1072 (2008)

Table 3. The results of stage-1 (query categorization) (left:3-class,right:4-class)

query	entity	Labeled Class	
Gibson	Gibson	1	1
Lexus	Lexus	1	1
McDonald's	McDonald's	1	1
sony	sony	1	1
Starbucks	Starbucks	1	1
apache	apache	1	1
oracle	Oracle	1	1
friday's	friday's	1	3
Amazon	Amazon.com	3	3
Blizzard	Blizzard Entertainment	3	3
fox	fox channel	3	3
jaguar	Jaguar Cars Ltd.	3	3
muse	muse band	3	3
sharp	Sharp Corporation	3	3
Apple	Apple	3	3
seat	seat	3	3
subway	subway	3	3
Cisco	Cisco Systems	3	4
ford	Ford Motor Company	3	4
McLaren	McLaren Group	3	4
stanford	Stanford Junior University	3	4
Yale	Yale University	3	4
canon	Canon inc.	3	4
CVS	CVS/pharmacy	3	4
emory	Emory University	3	4
GM	General Motors	3	4
MTV	MTV	3	4
Orange	Orange	2	3
scorpions	scorpions	2	3
sonic	sonic.net	2	3
tesla	Tesla Motors	2	3
johnnie	Johnnie Walker	2	3
Liverpool	Liverpool FC	2	3
mac	macintosh	2	4
camel	camel	2	2
Denver	Denver Nuggets	2	2
Deutsche	Deutsche Bank	2	2
kiss	kiss band	2	2
jfk John	F. Kennedy International Airport	2	2
Lloyd	Lloyds Banking Group	2	2
Metro	Metro supermarket	2	2
Milan	A.C. Milan	2	2
Paramount	Paramount Group	2	2
Roma	A.S. Roma	2	2
US	US Airways	2	2
Virgin	Virgin Media	2	2
zoo	Zoo Entertainment	2	2

Table 4. Results of labeling the queries in training set (left:3-class,right:4-class)

query	entity	Labeled Class	
nikon	nikon	1	1
linux	linux	1	1
Lufthansa	lufthansa	1	1
Foxtel	foxtel	1	1
alcatel	alcatel	1	1
Renault	renault	1	1
lamborghini	lamborghini	1	1
Yamaha	yamaha	1	1
Fujitsu	fujitsu	1	1
Marriott	marriott international	1	1
Marvel	marvel comics	1	3
philips	philips	1	3
Mercedes	mercedes-benz	1	3
Mandalay	mandalay bay resort and casino	1	3
armani	armani	1	3
barclays	barclays	1	3
Blockbuster	blockbuster inc.	1	3
bayer	bayer	3	3
fender	fender	3	3
cadillac	cadillac	3	3
Rover	land rover	3	3
BART	bart	3	4
Luxor	luxor hotel and casino	3	4
Boingo	boingo (wifi for travelers)	3	4
MGM	mgm grand hotel and casino	3	4
Harpers	harpers	3	4
Edmunds	edmunds.com	3	4
MTA	mta bike plus (nyc)	3	4
Southwest	southwest airlines	2	4
dunlop	dunlop	2	4
Amadeus	amadeus it group	2	4
pioneer	pioner company	2	2
Magnum	magnum research	2	2
mdm	mdm (event agency)	2	2
MEP	mep	2	2
Mercer	mercer consulting	2	2
Impulse	impulse (records)	2	2
elf	elf corporation	2	2
Apollo	apollo hospitals	2	2
Craft	craft magazine	2	2
nordic	nordic airways	2	2
Emperor	emperor entertainment group	2	2
folio	folio corporation	2	2
Smarter	smarter travel	2	2
Liquid	liquid entertainment	2	2
Lynx	lynx express	2	2
bulldog	bulldog solutions	2	2
shin	shin corporation	2	2
pierce	pierce manufacturing	2	2
Renaissance	renaissance technologies	2	2
Mack	mack group	2	2
Delta	delta holding	2	2