JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 10, No. 4, December 2016, pp. 128-136

Company Name Discrimination in Tweets using Topic Signatures Extracted from News Corpus

Beomseok Hong and Yanggon Kim
Department of Computer and Information Science, Towson University, Towson, MD, USA
bhong1@students.towson.edu, ykim@towson.edu

Sang Ho Lee*
School of Software, Soongsil University, Seoul, Korea
shlee199@gmail.com

Abstract

It is impossible for any human being to analyze the more than 500 million tweets that are generated per day. Lexical ambiguities on Twitter make it difficult to retrieve the desired data and relevant topics. Most of the solutions for the word sense disambiguation problem rely on knowledge base systems. Unfortunately, it is expensive and time-consuming to manually create a knowledge base system, resulting in a knowledge acquisition bottleneck. To solve the knowledge-acquisition bottleneck, a topic signature is used to disambiguate words. In this paper, we evaluate the effectiveness of various features of newspapers on the topic signature extraction for word sense discrimination in tweets. Based on our results, topic signatures obtained from a snippet feature exhibit higher accuracy in discriminating company names than those from the article body. We conclude that topic signatures extracted from news articles improve the accuracy of word sense discrimination in the automated analysis of tweets.

Category: Human computing

Keywords: Twitter; Tweet; Word sense discrimination; Topic signature

I. INTRODUCTION

Social media has significantly impacted our lives and has changed the way people communicate with one another. People share their feelings, emotions, and opinions by posting short texts, images, and videos on social media sites such as Twitter, Facebook, and Instagram. The usage of social media is increasing in tandem with the rapid growth of smartphone ownership. Marketers therefore pay attention to social media for advertising their products, and social media has become an influential viral marketing tool [1]. They are able to delve into immense amount of social media data to examine the popularity, reputation, and people's opinions of their products and brands [2]. Twitter, one of the most popular social media platforms, serves as an electronic word-of-mouth (eWOM) that affects customer's buying decisions by sharing opinions and information about brands [3]. According to Twitter Inc., there are 320 million monthly active users, and there are 1 billion unique visits to sites with embedded tweets per month. Over 500 million tweets were generated every day as of 2015. Due to the

Open Access http://dx.doi.org/10.5626/JCSE.2016.10.4.128 http://jcse.kiise.org

Received 13 November 2016; **Accepted** 05 December 2016
*Corresponding Author

pISSN: 1976-4677 eISSN: 2093-8020

enormous amount of tweets, it is impossible for a human to analyze them. Furthermore, the shortness and informality of tweets, including grammatical errors, misspellings, and unreliable capitalizations increase the difficulty of understanding tweets.

In addition, lexical ambiguity distracts from relevant topics and is a pervasive problem inhibiting researchers' abilities to retrieve desired data [4]. Many words can be interpreted in multiple ways. Unlike humans, machines need a process to understand the underlying meaning of the words. Word Sense Disambiguation (WSD) is a way for machines to understand the meaning of words, and different possible solutions have been suggested over decades [5]. Most of the solutions rely on knowledge-based approaches such as thesauri, ontologies, or sense-annotated corpora. Unfortunately, it is expensive and time-consuming to create the knowledge base system manually, and this problem is called the knowledge acquisition bottleneck [6]. A topic signature is used to solve the knowledge acquisition bottleneck. The topic signature is a family of words related to a given topic, and it is used for summarizing a document in the early stages. Manually-annotated knowledge base systems such as WordNet has insufficient lexical and semantic information. Therefore topic signatures from large-scale resources perform better than manually-annotated knowledge base systems [7].

Newspapers are used in the field of WSD as external knowledge resources. Since collections of newspapers are unstructured resources, they need to be annotated with senses [8], or used as raw corpora. Recently, The New York Times, an American daily newspaper, has provided public access to their newspaper repository using application programming interfaces (APIs). Newspapers can be retrieved by sending a query term, and the search engine retrieves news articles relevant to the query term. Consequently, only relevant newspapers can be retrieved from a large-scale repository. Together with other information retrieval techniques, we can collect the corpus related to a target sense without manual annotations. Moreover, new words that reflect certain topics will arise over time. The new words can be discovered as topic signatures by accessing up-to-date newspapers in the repository using The New York Times APIs.

In previous research studies, we proposed a system for discriminating company names in Twitter based on knowledge from news content without manual annotations [9]. Despite the fact that the New York Times APIs provide a variety of features such as abstracts, headlines, lead paragraphs, and snippets, the previous system utilized only the main text of the newspaper (i.e., body of each article) for automatic word sense discrimination.

In this paper, we evaluate the effectiveness of the various features of newspapers on the use of topic signature extraction for word sense discrimination. We initially exploit a retrieval rate to evaluate the usefulness of the features, and determine thresholds for topic signature extraction. Lastly, we evaluate the accuracy and f-measure to examine the most useful feature for company name discrimination. This paper is organized as follows: in Section II we provide the background to WSD, topic signatures, and related research. In Section III we present the method to build topic signatures from news articles. Additionally, the experimental setup and result are described in Sections IV and V. Finding and limitations are explained with the concluding remarks in Section VI.

II. RELATED WORK

A. Word Sense Disambiguation

WSD has been one of the research areas in Natural Language Processing (NLP) for several decades. The word sense disambiguation is generally divided into supervised, unsupervised approaches.

Supervised approaches use various machine learning methods with manually annotated resources for identifying word senses. Various supervised methods have been adopted such as a Nave Bayes, neural network, instance-based learning, support vector machine, and ensemble methods. Building manually annotated resources is an expensive and time-consuming work as documents and data on the Web grow continuously. In an effort to resolve the knowledge acquisition bottleneck, a bootstrapping method and a topic signature are adopted to the word sense disambiguation [10, 11]. In SemEval-2007, an international word sense disambiguation competition, the best system achieved an 88.70% accuracy whereas the first sense baseline achieved 78% [12]. A gold standard data constructed by the manually tagged Wall Street Journal corpus was used for the evaluation. The accuracy of the disambiguation is relatively high because a newspaper corpus is a long document and contains enough clue words to be used for the disambiguation.

Recently, many studies have tried to discriminate word senses in Twitter. Due to the fact that tweets are usually short and informal, it is much more difficult for machines to understand the word senses of tweets. The third Web People Search (WePS-3) task-2 evaluation campaign was held to address the ambiguity of named entities in Twitter and to encourage research groups to resolve the problem by providing the information of companies and collections of tweets for each company. Several groups participated in the competition [13]. The best system is LSIR-EPFL which builds six profiles of each company from external sources such as the home page, the metadata of the website, the category profile using WordNet, GoogleSet, and the manually user-defined terms for both positive and negative aspects [14]. The second best system ITC-UT makes use of six rules to categorize a company bias on tweets into 3 or 4 classes. For each bias, a procedure of the tweet classification is differently specified [15].

Most of the research tries to resolve the named entity ambiguity using external sources such as Wikipedia, Google search, DBpedia, and the company's home page. None of the research has exploited the news corpus for the tweet discrimination.

B. Topic Signatures

A topic signature is a family of terms that are highly correlated with a target concept and is defined as follows:

$$TS = \{topic, signature\} \\ = \{topic, < (t_1, w_1), ..., (t_n, w_n >\} \quad (1)$$

where *topic* is the target concept and *signature* is a vector of related terms [16]. Each t_i is a term highly correlated to *topic* with a weight w_i. A topic signature is a statistical approach that exploits the natural tendency of the semantically related words which co-occur more often than by chance in the same context [17]. Topic signatures are typically extracted from a pre-classified corpus because the relatedness of topic signatures is generally measured by tf-idf, the chi-squared test, or mutual information.

Despite the weighting of the methods, irrelevant words can be assigned high weights in topic signatures. Although filtering the irrelevant words outwardly refines the topic signatures, it does not have much effect on performance [18].

III. A COMPANY NAME DISCRIMINATION IN TWEETS

Company name discrimination is considered a binary classification task that checks whether the classification result would be related or non-related to a target company. Fig. 1 depicts the overall process of the proposed approach for the company name discrimination in tweets. Firstly, the New York Times (NYT) articles related to a target company are collected from the NYT repository using APIs.

Various features are available in the articles, such as abstracts, headlines, lead paragraphs, snippets and article bodies. Before extracting topic signatures, news articles are converted into the bag-of-words representation. In the bag-of-words model, a document is represented as a bag of words and the word order is ignored. Texts are segmented for the bag-of-words representation and tagged with their parts-of-speech. Only nouns are extracted for the company name discrimination because most nouns are concrete and used for the subject and object of the sentence. After NLP processing, topic signatures are extracted from the collected news articles by the document frequency. A straightforward classification method classifies the tweets based on the topic signatures. The classification method determines whether a tweet is relevant or irrelevant to a given company. In this section we explain how to collect news articles from the NYT repository, the topic signature extraction process, and the classification method in detail.

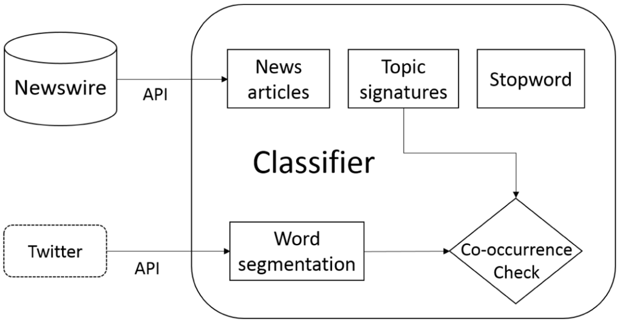

Fig. 1. An overview of the system architecture.

A. Collecting News Articles

News articles can be collected by a search keyword using the NYT APIs up to 1,010 articles per search keyword. The search keyword should be carefully decided to retrieve relevant articles. Searching for a full name of the company mitigates the ambiguity in the collected articles by retrieving more related articles so that the articles can be used as an external knowledge resource without manual annotations. Although we rely on the NYT APIs to collect relevant articles, the verification of the collected articles is necessary. After examining the collected articles, we observed that the full name of the company (e.g. Apple Inc.) retrieves related articles much more than the part of the company name (e.g. apple), and the collected articles can be used as an external knowledge resource.

Various features are available in each collected article and the four features associated with contents are suitable for extracting topic signatures: bodies, abstracts, lead paragraphs, and snippets. The body feature is a full text of the news article and the other features are the short summaries of the news article.

However, detailed explanations of the features are not specified in the NYT API document.

B. Topic Signature Extraction

Although various features are available in the news articles, it is necessary to examine which feature is more useful for the company name discrimination. Some of data are empty and null data can be retrieved from the NYT repository. Retrieval rate is used to determine how much data in the feature is available in the collected articles, and it is defined as:

$$\text{Retrieval rate} = \frac{\textit{Non-empty data in the feature}}{\textit{Total collected articles}} \qquad (2)$$

A higher retrieval rate implies the feature has less missing data and more available data in the collected articles. Consequently, a feature with a higher retrieval rate is more likely to be useful for extracting topic signatures.

A topic signature is a word vector that is topically related to a target word sense. Since the topic signature is the key to discriminate word senses, it is important to extract meaningful topic signatures from news articles. Typically, topic signatures are extracted by comparing the occurrence in related articles and unrelated articles to a target word sense.

While related articles can be collected by the search engine, it is nearly impossible to construct unrelated articles without human annotations. For this reason topic signatures are extracted from only related articles.

Topic signatures are extracted from the high retrieval rate features based on the document frequency. The document frequency of a word is defined as the number of articles that contain the word in the collected articles. Since one news article usually covers one topic, the document frequency is the main criterion to discover the topic signatures. Vocabulary is a list of unique words in the collected articles. Topic signatures are extracted from the vocabulary of each feature if the words document frequency is greater than the threshold which is heuristically determined.

C. Classification

The classification method uses the occurrence of topic signatures in tweets because tweets are short and each word in a tweet has a strong meaning. Algorithm 1 explains the algorithm of the tweet classification. Topic signatures are extracted from news articles, and tweets and topic signatures are prepared in the form of the bag-of-words model. In the classification, the tweet is classified to a related tweet to a target company if any topic signatures occur in the tweet. If no topic signatures occur in the tweet, the tweet is classified to a non-related tweet to a target company.

Algorithm 1. Tweet classification

```
tweet ← $(t_1, t_2, t_3, ..., t_n)$
topic signature ← $(w_1, w_2, w_3, ..., w_n)$
procedure RELATED(tweet,topic signature)
    if (tweet ∩ topic signature) ≠ ∅ then
        set tweet related
    else
        set tweet non-related
    end if
end procedure
```

Table 1. A list of 27 companies used in the evaluation

27 companies (full names used in news search)
Amazon.com, Apache Software Foundation, Apple Inc., Blizzard Entertainment, Canon Inc., Cisco Systems, CVS/pharmacy, Ford Motor Company, T.G.I. Friday's, General Motors, Gibson Guitar, Jaguar Cars Ltd., Lexus, McDonald's, Metro Supermarket, Oracle Corporation, Orange S.A., Paramount Group, Seat S.A., Sharp Corporation, Sonic.net, Sony, Starbucks, Subway, Tesla Motors, US Airways, Virgin Media

IV. EXPERIMENTAL SETUP

We now describe metrics for the evaluation, data sets and data preprocessing for the experiment.

A. Evaluation Metrics

For the performance evaluation, we estimated the accuracy, precision, recall and f-measure for each company. The measures are defined as:

$$\text{Accuracy} = \frac{TP+TN}{N} \qquad (3)$$

$$\text{Precision (related)} = \frac{TP}{TP+FP} \qquad (4)$$

$$\text{Recall (related)} = \frac{TP}{TP+FN} \qquad (5)$$

$$\text{Precision (non-related)} = \frac{TN}{TN+FN} \qquad (6)$$

$$\text{Recall (non-related)} = \frac{TN}{TN+FP} \qquad (7)$$

$$\text{F-measure} = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (8)$$

where N is the number of tweets, TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

B. Data Sets and Preprocessing

In WePS-3 task 2 Online Reputation Management, 47 named entities and around 500 tweets corresponding to each named entity are provided as a test set.

The test set is labelled by 5 human annotators using Amazon Mechanical Turk. We categorized the 47 provided named entities and selected only 27 company-related entities for the experiment since companies are occasionally main topics of newspapers and company names are frequently mentioned in newspapers. The experiment was carried out with the 27 companies listed in Table 1. The total number of tweets provided by

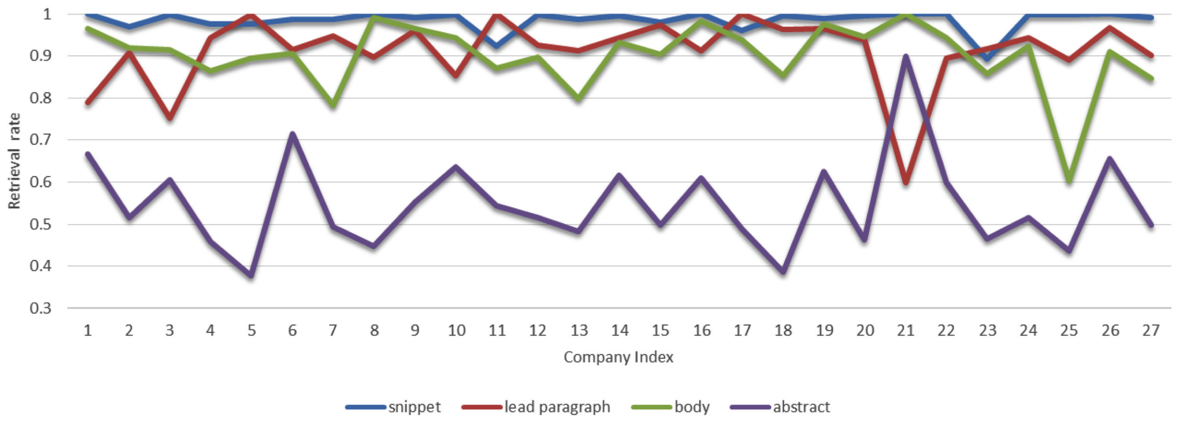

Fig. 2. The retrieval rate of news article features.

WePS-3 for 27 companies is 11,526 tweets. In this experiment the URL addresses and username tags (@) were ignored and the hash tags (#) were regarded as normal words.

The total number of news articles we collected for 27 companies is 20,492 articles. We collected the body, the abstract, the lead paragraph, and the snippet from each article. All the raw texts were converted to a bag-of-words representation using word segmentation and the part-of-speech tagger in the Stanford Natural Language Processing (NLP) module. Only nouns are extracted for the company name discrimination because most nouns are concrete and used for the subject and object of the sentence.

In the preliminary experiment, nouns showed the best accuracy in the tweet classification compared to other parts of speech.

V. EXPERIMENTAL RESULT

We now describe the results of the feature selection, threshold for the topic signature extraction, and the evaluation of the company name discrimination.

A. Feature Selection

Fig. 2 illustrates the retrieval rate of the four different features of news articles. The snippet feature obtained the 0.98 retrieval rate on average, which is the highest retrieval rate in the four features. A feature with a higher retrieval rate is more likely to be useful for extracting topic signatures. For example, when we collect 1,000 news articles, 980 news articles contain the snippets, and 20 news articles do not have the snippets. The average retrieval rate of the lead paragraph feature is 0.91, and that of the body feature is 0.90. The abstract feature obtained a 0.54 retrieval rate, which is relatively low. Despite the high retrieval rate, the lead paragraph feature is excluded from the experiment. The contents of the lead paragraph feature are almost identical to those of the snippet feature but the retrieval rate of the snippet feature is higher. According to the average retrieval rate, we selected the snippet feature and the body feature as candidate features for extracting topic signatures. The main difference between the snippet feature and the body feature is their respective lengths. The snippet consists of at most 2 sentences whereas the body contains several paragraphs.

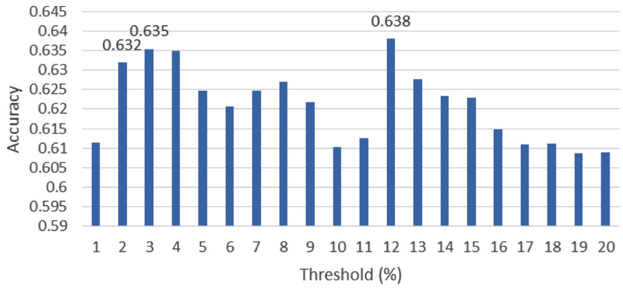

Fig. 3. The accuracy of the classification result on various thresholds.

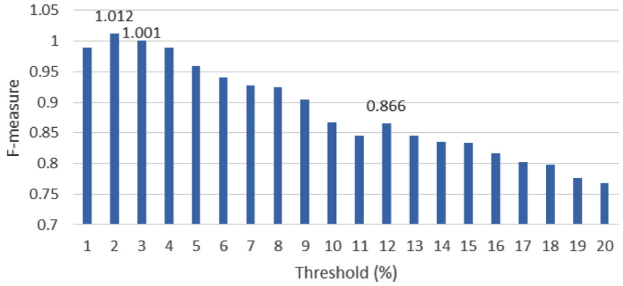

Fig. 4. The f-measure of the classification result on various thresholds.

B. Threshold for Extracting Topic Signatures

Topic signatures are extracted based on a threshold of

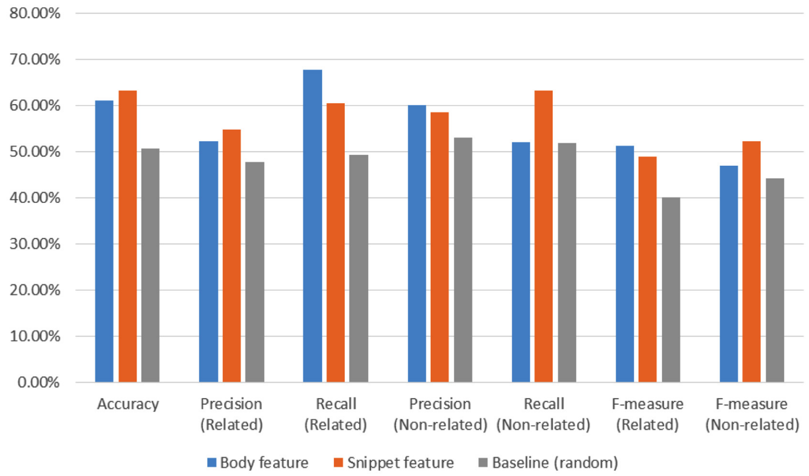

Fig. 5. A comparison of the evaluation result.

Table 2. The average number of words in topic signatures for each threshold

Threshold (%)	Average topic signatures (words)
2	75.81
3	44.11
12	6.07

the document frequency. It is important to find the reasonable threshold to exclude insignificant words from the topic signature. If the threshold is too low, superfluous words would be included. On the other hand, if the threshold is too high, it leads to a false bias and most of the results are labelled as non-related since too few words would be included in the topic signature. Because the number of the collected news articles is different for each company, we set the threshold by the ratio of collected documents.

For example, if the threshold is set to 15%, the words whose document frequency is greater than N*0.15 are chosen as a topic signature where N is the total number of the news articles for the given company. The threshold for the body feature is already determined at 15% of the total number of the news articles in the previous research [9].

In the same manner, we compared various thresholds for the snippet feature to determine the best threshold. As shown in Figs. 3 and 4, various thresholds from 1% to 20% were tested. In Fig. 3 the accuracy is fluctuating and unpredictable as the threshold changes. The 12% threshold achieved the highest accuracy. Fig. 4 shows the sum of the f-measure of the related class and the non-related class.

As the threshold increases, the f-measure has fallen off steadily. The 2% threshold achieved the highest f-measure. We compared the number of extracted words in topic signatures to figure out what value constituted a reasonable threshold. As shown in Table 2, on average, the 2% threshold extracted 75.81 words as topic signatures; the 3% threshold extracted 44.11 words, and 12% threshold extracted 6.07 words. As a result, the 2% threshold has the most topic signatures without decreasing the performance in both accuracy and f-measure. Therefore, the threshold for the snippet feature is determined as 2% of the total number of news articles.

C. Evaluation

The performance of the classification is measured by accuracy, precision, recall, and the f-measure. A random baseline is used to evaluate the improvement of company name classifications in tweets. Fig. 5 shows an evaluation of the result compared to the baseline.

A random baseline consists of all the randomly labeled tweets. The result shows that the topic signatures extracted from the article body increased the accuracy by 10.4% and those of the snippet feature increased the accuracy by 12.5% as compared with the random baseline. The precision, recall, and f-measure for both snippet and body feature are also increased when compared with the random baseline. Word sense discrimination was more accurate when using a snippet feature compared to using the body of the article. Whereas f-measure of the body feature in the related class is higher than that of the snippet feature, the f-measure of the body feature in the non-related class is lower than that of the snippet feature. We speculate the reason for the difference of the f-measures between the related class and non-related class is that the body feature has more topic signatures (140.44 words on average) than the snippet feature (75.81 words on average).

Tables 3 and 4 illustrate the results of the top 5 out of the 27 companies in order of the sum of both related and non-related f-measures for the body feature and the snippet feature respectively.

Table 3. The top 5 companies using the body feature by f-measure

Entity	Accuracy (%)	Related			Non-related		
		Precision	Recall	F-measure	Precision	Recall	F-measure
Apple Inc.	83.5	0.942	0.856	0.897	0.495	0.729	0.590
General Motors	74.2	0.659	0.872	0.751	0.861	0.638	0.733
Apache Software	71.0	0.677	0.751	0.712	0.748	0.674	0.709
Tesla Motors	69.3	0.497	0.858	0.630	0.909	0.621	0.738
Jaguar Cars Ltd.	66.3	0.683	0.728	0.705	0.635	0.583	0.608

Table 4. The top 5 companies using the snippet feature by f-measure

Entity	Accuracy (%)	Related			Non-related		
		Precision	Recall	F-measure	Precision	Recall	F-measure
General Motors	81.1	0.876	0.672	0.761	0.778	0.924	0.844
Tesla Motors	77.0	0.585	0.834	0.688	0.911	0.742	0.818
Jaguar Cars Ltd.	68.1	0.766	0.609	0.678	0.615	0.771	0.684
Apple Inc.	73.0	0.948	0.717	0.816	0.353	0.797	0.489
Gibson Guitar	71.4	0.347	0.717	0.468	0.922	0.713	0.804

Companies with high scores on f-measures indicate that they are well discriminated in our approach for both related and non-related tweets. We observed that Apple Inc., General Motors, Tesla Motors, and Jaguar Cars Ltd. are highly ranked in terms of accuracy and f-measure for both using the body feature and the snippet feature because product-related words are extracted from news articles as topic signatures. For example, topic signatures of automobile manufacturers include car-related words such as vehicle, car, model, etc. Topic signatures of Apple Inc. include names of their products such as iPhone, iPod, etc.

VI. DISCUSSION AND CLOSING REMARKS

For natural language processing, we exploit the Stanford Natural Language Processing (NLP) module. The Stanford NLP performs well on the news article. However, the informality of tweets such as grammatical errors, misspellings, and unreliable capitalizations may depreciate the quality of the NLP module in tweets. For example, a tweet This Is Apples Next iPhone (http://bit.ly/cEJuUq) is segmented into be, apple, next, bit.ly, cejuuq through the Stanford NLP because of the lack of the blank space in between the words. The word iPhone, a topic signature of Apple Inc., is not extracted from the tweet due to both the limitation of the NLP module and the informality of the tweet.

A time gap between the collected news articles and the evaluation data may influence the classification result. When we collect news articles, the latest articles are preferentially retrieved by the New York Times APIs. For example, out of 1,010 news articles about Apple Inc., 299 news articles were generated after January 2015. As privacy became a controversial issue for Apple Inc. in recent years, the word privacy is selected as a topic signature as shown in Table 5. On the other hand, the evaluation data was released in 2010. This time gap may affect the evaluation result.

Table 5. A topic signature for Apple Inc.

Topic signature (Apple Inc.)
watch, tech, computer, company, executive, tablet, mac, government, steve, operating, smartphone, ios, music, version, system, share, application, app, case, phone, market, iPad, user, customer, steven, jobs, iPod, device, privacy, technology, software, problem, iPhone, feature, cook, security

We propose a semi-supervised system for a company name discrimination on tweets based on topic signatures extracted from news articles. The proposed system is a fully automated system that requires only a search keyword when adding a new company. From the experiment we found that news articles could be used to disambiguate tweets as an external source. However, we have not deeply analyzed the various features of news articles in the previous research. In this paper we conduct an experiment for measuring the effectiveness of various features in news articles for the company name discrimination. The snippet, lead paragraph, and body feature obtain high retrieval rates, meaning that they are useful features for extracting topic signatures. However, only the snippet and body feature are selected as candidate features because almost every lead paragraph has the same contents as the snippet does. The best threshold for extracting the topic signature is determined as 2% for the snippet feature and 15% for the body feature. The classification result for each feature is 63.2% accuracy for the snippet feature and 61.1% for the body feature. As compared with the random baseline, the accuracy is increased by 10.4% with the body feature and 12.5% with the snippet feature. Although the body feature extracted topic signature words twice as much as those of the snippet feature, the snippet feature achieved 2.1% higher accuracy than that of the body feature. We conclude that topic signatures extracted from news articles improve the accuracy of the company name discrimination in Twitter.

REFERENCES

1. R. K. Miller and K. Washington, *The 2013 Entertainment, Media & Advertising Market Research Handbook*, 13th ed., Loganville, GA: Richard K Miller & Associates, 2013
2. W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: a case study in the pizza industry," *International Journal of Information Management,* vol. 33, no. 3, pp. 464-472, 2013.
3. B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: tweets as electronic word of mouth," *Journal of the American Society for Information Science and Technology,* vol. 60, no. 11, pp. 2169-2188, 2009.
4. R. Krovetz and W. B. Croft, "Lexical ambiguity and information retrieval," *ACM Transactions on Information Systems*, vol. 10, no. 2, pp. 115-141, 1992.
5. R. Navigli, "Word sense disambiguation: a survey," *ACM Computing Surveys,* vol. 41, no. 2, pp. 1-69, 2009.
6. W. A. Gale, K. W. Church, and D. Yarowsky, "A method for disambiguating word senses in a large corpus," *Computers and the Humanities,* vol. 26, no. 5/6, pp. 415-439, 1992.
7. M. Cuadros and G. Rigau, "Quality assessment of large scale knowledge resources," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing,* Sydney, Australia, 2006, pp. 534-541.
8. S. Landes, C. Leacock, and R. I. Tengi, "Building semantic concordances," in *WordNet: An Electronic Lexical Database*, Cambridge, MA: MIT Press, 1998, pp. 199-216.
9. B. Hong, Y. Han, and Y. Kim. "A semi-supervised tweet classification method using news articles," in *Proceedings of the 2015 Conference on Research in Adaptive and Convergent Systems*, Prague, Czech Republic, 2015, pp. 62-67.
10. R. Mihalcea, "Co-training and self-training for word sense disambiguation," in *Proceedings of the 8th Conference on Computational Natural Language Learning (CoNLL),* Boston, MA, 2004, pp. 33-40.
11. E. Agirre, O. Ansa, D. Martinez, and E. Hovy, "Enriching WordNet concepts with topic signatures," in *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources,* Pittsburgh, PA, 2001, pp. 23-28.
12. E. Agirre and A. Soroa, "SemEval-2007 task 02: evaluating word sense induction and discrimination systems," in *Proceedings of the 4th International Workshop on Semantic Evaluations*, Prague, Czech Republic, 2007, pp. 7-12.
13. E. Amigo, J. Artiles, J. Gonzalo, D. Spina, B. Liu, and A. Corujo, "WePS-3 evaluation campaign: overview of the online reputation management task," in *Proceedings of International Conference on Cross-Language Evaluation Forum (CLEF2010)*, Padua, Italy, 2010.
14. S. R. Yerva, Z. Miklos, and K. Aberer, "It was easy, when apples and blackberries were only fruits," in *Proceedings of International Conference on Cross-Language Evaluation Forum (CLEF2010)*, Padua, Italy, 2010.
15. M. Yoshida, S. Matsushima, S. Ono, I. Sato, and H. Nakagawa, "ITC-UT: tweet categorization by query categorization for on-line reputation management," in *Proceedings of International Conference on Cross-Language Evaluation Forum (CLEF2010)*, Padua, Italy, 2010.
16. C. Y. Lin and E. Hovy, "The automated acquisition of topic signatures for text summarization," in *Proceedings of the 18th Conference on Computational Linguistics,* Saarbrucken, Germany, 2000, pp. 495-501.
17. M. Biryukov, R. Angheluta, and M. F. Moens, "Multidocument question answering text summarization using topic signatures," *Journal of Digital Information Management,* vol. 3, no. 1, pp. 27-33, 2005.
18. E. Agirre and O. L. de Lacalle, "Publicly available topic signatures for all WordNet nominal senses," in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 2004, pp. 1123-1126.

Beomseok Hong

Beomseok Hong is currently a PhD candidate in the Department of Computer and Information Sciences at Towson University. His research interests include text mining, information retrieval, and question answering.

Yanggon Kim

Yanggon Kim received his B.S. and M.S. degrees from Seoul National University, Seoul, Korea in 1984 and 1986, respectively, and Ph.D. degree in Computer Science and Engineering from Pennsylvania State University, Pennsylvania, 1995. He is currently a Professor in the Department of Computer and Information Sciences, Towson University, Maryland. His current research interests include computer networks, secure BGP network, distributed computing systems, big data and data sciences in social networks.

Sang Ho Lee

Sang Ho Lee received his B.S. degree from Seoul National University, Seoul, Korea in 1984, and M.S. and Ph.D. degrees in Computer Science from Northwestern University, Illinois in 1986 and 1989, respectively. He is currently a Professor in the School of Software, Soongsil University, Seoul, Korea. He served as president of Korea Association of University Information and Computer (KAUIC), and Korea Private University Library Association during early 2010s. His research interests include database systems, Web databases, and social computing.