A Generative Model for Identifying Target Companies of Microblogs

Yeyun Gong, Yaqian Zhou, Ya Guo, Qi Zhang, Xuanjing Huang
Shanghai Key Laboratory of Intelligent Information Processing
School of Computer Science, Fudan University
825 Zhangheng Road, Shanghai, P.R.China
{12110240006, zhouyaqian, 13210240002, qz, xjhuang}@fudan.edu.cn

Abstract

Microblogging services have attracted hundreds of millions of users to publish their status, ideas and thoughts, everyday. These microblog posts have also become one of the most attractive and valuable resources for applications in different areas. The task of identifying the main targets of microblogs is an important and essential step for these applications. In this paper, to achieve this task, we propose a novel method which converts the target company identification problem to the translation process from content to targets. We introduce a topic-specific generative method to model the translation process. Topic specific trigger words are used to bridge the vocabulary gap between the words in microblogs and targets. We examine the effectiveness of our approach via datasets gathered from real world microblogs. Experimental results demonstrate a 20.2% improvement in terms of F1-score over the state-of-the-art discriminative method.

1 Introduction

With the rapid growth of social media, about 72% of adult internet users are also members of a social networking site[1]. Over the past few years, microblogging has become one of the most popular services. Meanwhile, microblogs have also been widely used as sources for analyzing public opinions (Bermingham and Smeaton, 2010; Jiang et al., 2011), prediction (Asur and Huberman, 2010; Bollen et al., 2011), reputation management (Pang and Lee, 2008; Otsuka et al., 2012), and many other applications (Bian et al., 2008; Sakaki et al., 2010; Becker et al., 2010; Guy et al., 2010; Lee and Croft, 2013; Guy et al., 2013). For most of these applications, identifying the microblogs that are relevant to the targets of interest is one of the basic steps (Lin and He, 2009; Amigó et al., 2010; Qiu et al., 2011; Liu et al., 2013). Let us firstly consider the following example:

Example 1: *11" MacBook Air can run for up to five hours on a single charge.*

"*MacBook Air*" can be considered to be the target being discussed on the microblog, and we can also infer from the microblog that it is related to Apple Inc. The ability to discriminate which company is being referred to in a microblog is required by many applications.

Previous studies on fine-grained sentiment analysis and aspect-based opinion mining proposed supervised (Popescu and Etzioni, 2005; Liu et al., 2012a; Liu et al., 2013) and unsupervised methods (Hu and Liu, 2004; Wu et al., 2009; Zhang et al., 2010) to extract targets of opinion expressions. Based on the associations between opinion targets and opinion words, some methods were also introduced to simultaneously solve the opinion expression and target extraction problems (Qiu et al., 2011; Liu et al., 2012a). However, most of the existing methods in this area only focus on extracting items about which opinions are expressed in a given domain. The implicated information of targets is rarely considered. Moreover, domain adaptation is another big challenge for these fine-grained methods in processing different domains.

[1] It is reported by the Pew Research Center's Internet & American Life Project in Aug 5, 2013.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 688–697, Dublin, Ireland, August 23-29 2014.

The WePS-3[2] (Amigó et al., 2010) and RepLab 2013[3] (Amigó et al., 2013) evaluation campaigns also addressed the problem from the perspective of the disambiguation of company names in microblogs. Microblogs that contain company names at a lexical level are classified based on whether it refers to the company or not. Various approaches have been proposed to address the task with different methods (Pedersen et al., 2006; Yerva et al., 2010; Zhang et al., 2012; Spina et al., 2012; Spina et al., 2013). However, the microblogs that do not contain company names cannot be correctly processed using these methods. From analyzing the data, we observe that a variety of microblog posts belong to this type. They only contain products names, slang terms, and other related company content.

To achieve this task, in this paper, we propose the use of a translation based model to identify the targets of microblogs. We assume that the microblog posts and targets describe the same topic using different languages. Hence, the target identification problem can be regarded as a translation process from the content of the microblogs to the targets. We integrate latent topical information into the translation model to facilitate the translation process. Because product names, series, and other related information are important indicators for this task, we also incorporate this background knowledge into the model. To evaluate the proposed method, we collect a large number of microblogs and manually annotate a subset of these as golden standards. We compare the proposed method with state-of-the-art methods using the constructed dataset. Experimental results demonstrate that the proposed approach can achieve better performance than the other approaches.

2 The Proposed Method

2.1 The Generation Process

Given a corpus $D = \{d_i, 1 \leq i \leq |D|\}$, which contains a list of microblogs $\{d_i\}$. A microblog is a sequence of N_d words denoted by $w_d = \{w_{d1}, w_{d2}, ..., w_{dN_d}\}$. Each microblog contains a set of targets denoted by $c_d = \{c_{d1}, c_{d2}, ..., c_{dM_d}\}$. A word is defined as an item from a vocabulary with V distinct words indexed by $w = \{w_1, w_2, ..., w_V\}$. The nth word in the dth microblog is associated with not only one topic z_{dn}, but also an indicator variable l_{dn} which indicates whether w_{dn} belongs to the ontology ($l_{dn} = 1$), which contains company names, product names, series, and other related information, or is a common word ($l_{dn} = 0$). Each target is from the vocabulary with C distinct company names indexed by $c = \{c_1, c_2, ..., c_C\}$. The mth target in the dth microblog is associated with a topic z_{dm}. The notations used in this paper are summarized in Table 1. Fig. 1 shows the graphical representation of the generation process. The generative story for each microblog is as follows:

1. Sample word distribution $\phi^{t,l}$ from $Dir(\beta^l)$ for each topic $t = 1, 2, ..., T$ and each label $l = 1, ..., L$.

2. For each microblog d=1,2,...,$|D|$

   a. Sample topic distribution θ_d from $Dir(\alpha)$

   b. For each word $n = 1, 2, ..., N_d$

      i. Sample a topic $z_{dn} = t$ from $Multinomial(\theta_d)$

      ii. Sample a label $l_{dn} = l$ from the distribution over labels, $v^{d,n}$

      iii. Sample a word w according to multinomial distribution $P(w_{dn} = w | z_{dn} = t, l_{dn} = l, \phi^{t,l})$

   c. For each target $m = 1, 2, ..., M_d$

      i. Sample a topic $z_{dm} = t$ from $Multinomial(\theta_d)$

      ii. Sample a target $c_{dm} = c$ according to probability $P(c_{dm} = c | w_d, l_d, z_{dm} = t, B)$

As described above, we use l_{dn} to incorporate the ontology information into the model. In this work, we construct an ontology which contains 4,926 company names, 7,632 abbreviations, and 26,732 product names. These companies names are collected based on the top search queries in different categories [4]. We propose to use the distribution $v^{d,n}$ to indicate the probability of variable l_{dn}. We set $v^{d,n}$ by applying

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

[3] http://www.limosine-project.eu/events/replab2013

[4] http://top.baidu.com/boards

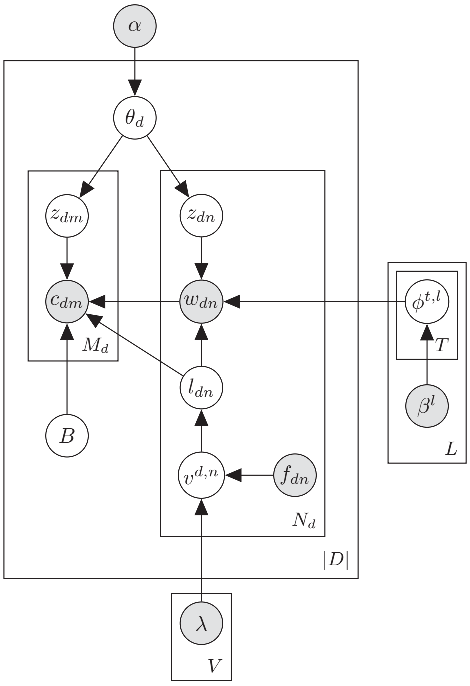

Figure 1: The graphical representation of the proposed model. Shaded circles are observations or constants. Unshaded ones are hidden variables.

various sources of ontology (presented by λ) and the context features of the word w_{dn} (presented by f_{dn}). In this work, we only consider the word itself as its context feature. This information is encoded into the hyperparameters $\{\lambda^w | w \in \{w_1, w_2, ..., w_V\}\}$, where λ^w is hyperparameter for the word w, and $\lambda_0^w + \lambda_1^w = 1$. For each word w in the ontology, we set λ_1^w to a value 0.9, λ_0^w to a value 0.1. For each word w not contained by ontology, we set λ_1^w to a value 0 and λ_0^w to a value 1. Based on the ontology, $v^{d,n}$ could be set as follows:

$$P(l_{dn} = l | w_{dn} = w) = v_l^{d,n} = \frac{\lambda_l^w}{\lambda_1^w + \lambda_0^w}, l \in \{0, 1\} \tag{1}$$

2.2 Model Inference

We use collapsed Gibbs sampling (Griffiths and Steyvers, 2004) to obtain samples of hidden variable assignment and to estimate the model parameters from these samples.

On the microblog content side, the conditional probability of a latent topic and label for the nth word in the dth microblog is:

$$Pr(z_{dn} = t, l_{dn} = l | w_{dn} = w, \mathbf{w}^{\neg n}, \mathbf{z}^{\neg n}, \mathbf{l}^{\neg n}) \propto \frac{\lambda_l^w}{\lambda_1^w + \lambda_0^w} \times \frac{N_{t,l}^{w,\neg n} + \beta^l}{N_{t,l}^{\neg n} + V\beta^l} \times \frac{N_d^{t,\neg n} + \alpha}{N_d^{\neg n} + T\alpha}, \tag{2}$$

where $N_{t,l}^{w,\neg n}$ is the number of the word w that are assigned to topic t under the label l; $N_{t,l}^{\neg n}$ is the number of all the words that are assigned to topic t under the label l; $N_d^{t,\neg n}$ is the number of topic t in the microblog d; $N_d^{\neg n}$ is the number of all the topics in the document d; $\neg n$ indicates taking no account of the current position n.

Given the conditional probability of $z_{dn} = t, l_{dn} = l$, we formalize the marginal probability of $z_{dn} = t$ as follows:

$$Pr(z_{dn} = t | w_{dn} = w, \mathbf{w}^{\neg n}, \mathbf{z}^{\neg n}, \mathbf{l}^{\neg n}) \propto \sum_{l=0}^{L-1} \frac{\lambda_l^w}{\lambda_1^w + \lambda_0^w} \times \frac{N_{t,l}^{w,\neg n} + \beta^l}{N_{t,l}^{\neg n} + V\beta^l} \times \frac{N_d^{t,\neg n} + \alpha}{N_d^{\neg n} + T\alpha} \tag{3}$$

Table 1: The notation used in the proposed model.

$\|D\|$	The number of microblogs in the data set
V	The number of unique words in the vocabulary
C	The number of companies
T	The number of topics
L	The number of labels
N_d	The number of words in the dth microblog
M_d	The number of companies in the dth microblog
w_d	All the words in the dth microblog
c_d	All the targets in the dth microblog
z_d	The topic of the words in the dth microblog
l_d	The label of the words in the dth microblog
B	The topic-specific word alignment table between a word and a target
$\phi^{t,l}$	Distribution of words for each topic t and each label l
θ_d	Distribution of topics in microblog d
$\upsilon^{d,n}$	Distribution of labels for word w_{dn}
$N_{t,l}^{w,\neg n}$	The number of the word w that is assigned to topic t under the label l except the position n
$N_{t,l}^{\neg n}$	The number of all the words that are assigned to topic t under the label l. except the position n
$N_d^{t,\neg n}$	The number of topic t in the microblog d except the position n
$N_d^{\neg n}$	The number of all the topics in the microblog d except the position n
$N_{t,l}^{c,w}$	The number of the target c that co-occurs with the word w labeled as l under topic t

After re-assigning the topic $z_{dn} = t$ for the current word, the conditional probability of ontology label for the nth word in the dth microblog is:

$$Pr(l_{dn} = l | w_{dn} = w, z_{dn} = t, \mathbf{w}^{\neg n}, \mathbf{z}^{\neg n}, \mathbf{l}^{\neg n}) \propto \frac{\lambda_l^w}{\lambda_1^w + \lambda_0^w} \times \frac{N_{t,l}^{w,\neg n} + \beta^l}{N_{t,l}^{\neg n} + V\beta^l} \tag{4}$$

On the target side, we perform topic assignments for each target as follows:

$$Pr(z_{dm} = t | c_{dm} = c, \mathbf{c}^{\neg m}, \mathbf{w}, \mathbf{l}, \mathbf{z}^{\neg m}) \propto \sum_{n=1}^{N_d} \delta^{l_{dn}} \frac{N_{t,l_{dn}}^{c,w_{dn},\neg m}}{N_{t,l_{dn}}^{w_{dn}} + \gamma C} \times \frac{N_d^{t,\neg m} + \alpha}{N_d^{\neg m} + T\alpha}, \tag{5}$$

where $\delta^{l_{dn}}$ is the weight for the label ($\delta^1 > 1, \delta^0 = 1$); $N_{t,l_{dn}}^{c,w_{dn},\neg m}$ is the number of the company c that co-occurs with the word w_{dn} labeled as l_{dn} under topic t; γC is a smoothing part; $N_{t,l_{dn}}^{w_{dn}}$ is the number of the word w_{dn} labeled as l_{dn} under topic t; $N_d^{t,\neg m}$ is the number of occurrences of topic t in the document d; $N_d^{\neg m}$ is the number of occurrences of all the topics in the document d; $\neg m$ indicates taking no account of the current position m.

Based on the above equations, after enough sampling iterations, we can estimate word alignment table B, $B_{c,w,t,l} = \delta^l \frac{N_{t,l}^{c,w}}{N_{t,l}^w + \gamma C}$. Some companies just occur few times, and most of the words co-occur with them also alignment with other companies, for this case, we use γC to smooth, where C represent the number of company c. And also we can estimate topic distribution θ for each document, and word distribution ϕ for each topic and each label, as follows:

$$\theta_d^t = \frac{N_d^t + \alpha}{N_d + T\alpha}, \quad \phi_w^{t,l} = \frac{N_{t,l}^w + \beta^l}{N_{t,l} + V\beta^l}$$

The possibility table $B_{c,w,t,l}$ has a potential size of $V \cdot C \cdot T \cdot L$. The data sparsity may pose a problem in estimating $B_{c,w,t,l}$. To reduce the data sparsity problem, we introduce the remedy in our model. We employ a linear interpolation with topic-free word alignment probability to avoid data sparsity problem:

$$B^*_{c,w,t,l} = \sigma B_{c,w,t,l} + (1-\sigma)P(c|w), \tag{6}$$

where $P(c|w)$ is topic-free word alignment probability between the word w and the company c. σ is trade-off of two probabilities ranging from 0.0 to 1.0.

2.3 Target Company Extraction

Just like standard LDA, the proposed method itself finds a set of topics but does not directly extract targets. Suppose we have a dataset which contains microblogs without targets, we can use the collapsed Gibbs sampling to estimate the topic and label for the words in each microblog. The process is the same as described in Section 3.2.

After the hidden topics and label of the words in each microblog become stable, we can estimate the distribution of topics for the dth microblog by: $P(t|w_d) = \theta^t_d = \frac{N^t_d+\alpha}{N_d+T\alpha}$. With the word alignment table B^*, we can rank companies for the dth microblog in unlabeled data by computing the scores:

$$Pr(c_{dm}|w_d) \propto \sum_{t=1}^{T}\sum_{n=1}^{N_d} P(c_{dm}|t, w_{dn}, l_{dn}, B^*) \cdot P(t|w_d)P(w_{dn}|w_d), \tag{7}$$

where $P(w_{dn}|w_d)$ is the weight of the word w_{dn} in the microblog content w_d. In this paper, we use inverse document frequency (IDF) score to estimate it. Based on the ranking scores calculated by Eq.(7), we can extract the top-ranked targets for each microblog to users.

3 Experiments

In this section, we will introduce the experimental results and datasets we constructed for training and evaluation. We will firstly describe the how we construct the datasets and their statistics. Then we will introduce the experiment configurations and baseline methods. Finally, the evaluation results and analysis will be given.

3.1 Datasets

We started by using Sina Weibo's API[5] to collect public microblogs from randomly selected users. The dataset contains 282.2M microblogs published by 1.1M users. We use *RAW-Weibo* to represent it in the following sections. Based on the collected raw microblogs, we constructed three datasets for evaluation and training.

3.1.1 Training data

Since social media users post thoughts, ideas, or status on various topics in social medias, there are a huge number of related companies. Manually constructing training data is a time consuming and cost process. In this work, we propose a weakly manual method based on ontology and hashtag. A hashtag is a string of characters preceded by the symbol #. In most cases, hashtags can be viewed as an indication to the context of the tweet or as the core idea expressed in the tweet. Hence, we can use hashtag as the targets.

We extract the microblogs whose hashtags contain ontology items as training data and the corresponding ontology items as targets. Obviously, the training data constructed based on this method is not perfect. However, since this method can effectively generate a great quantity of data, we think that general characteristics can be modeled with the generated training data. To evaluate the corpus, we randomly selected 100 microblogs from the training data and manually labeled their targets. The accuracy of the sampled dataset is 91%. It indicates that the proposed training data generation method is effective. From the *RAW-Weibo* dataset, we extracted a total of 1.79M microblogs whose hashtags contain more than one target. Training instances for 2,574 target companies are included in the training data.

[5] http://open.weibo.com/

3.1.2 Test data

For evaluation, we manually constructed a dataset *RAN-Weibo*, which contains 2,000 microblogs selected from *RAW-Weibo*. Three annotators were asked to label the target companies for each microblog. To evaluate the quality of annotated dataset, we validate the agreements of human annotations using Cohen's kappa coefficient. The average κ among all annotators is 0.626. It indicates that the annotations are reliable.

Since some targets are ambiguous, inspired by the evaluation campaigns WePS-3 and RepLab 2013, we also constructed a dataset *AMB-Weibo*, where microblogs include 10 popular company names which may cause ambiguity. For each target, we randomly selected and annotated 200 microblogs as golden standards. Three annotators were also asked to label whether the microblog is related the given target or not. The agreements of human annotations were also validated through Cohen's kappa coefficient. The average κ among all annotators is 0.692.

3.2 Experiment Configurations

We use precision (P), recall (R), and F1-score (F_1) to evaluate the performance. We ran our model with 500 iterations of Gibbs sampling. We use 5-fold cross-validation in the training data to optimize hyperparameters. The number of topics is set to 30. The other settings of hyperparameters are as follows: $\alpha = 50/T$, $\beta = 0.1$, $\delta = 20$, $\gamma = 0.5$. The smoothing parameter σ is set to 0.8.

For baselines, we compare the proposed model with the following baseline methods.

- **Naive Bayes (NB)**: The target identification task can be easily formalized as a classification task, where each target is considered as a classification label. Hence, we applied Naive Bayes to model the posterior probability of each target given a microblog.

- **Support Vector Machine (SVM)**: The content of microblogs are represented as vectors and SVM is used to model the classification problem.

- **IBM1**: Translation model (IBM model-1) is applied to obtain the alignment probability between words and targets.

- **TTM**: Topical translation model (TTM) was proposed by Ding et al. (2013) to achieve microblog hashtag suggestion task. We adopted it to estimate the alignment probability between words and targets.

3.3 Experimental Results

We evaluate the proposed method from the following perspectives: 1) comparing the proposed method with the state-of-the-art methods on the two evaluation datasets; 2) identifying the impacts of parameters.

Table 2 shows the comparisons of the proposed method with the state-of-the-arts discriminative and generative methods on the evaluation dataset *RAN-Weibo*. "*Our*" denotes the method proposed in previous sections. "*Our w/o BG*" represents the proposed method without background knowledge. From the results, we can observe that the proposed method is better than other methods. Discriminative methods achieve worse results than generative methods. We think that the large number of targets is one of the main reasons of the low performances. The results of the proposed models with and without ontology information also show that background knowledge can benefit both the precision and recall. TTM achieves better performance than IBM1. It indicates that topical information is useful for this task. The performances of our method are significantly better than TTM. It illustrates that our smoothing method and incorporation of background knowledge are effective.

From the description of the proposed model, we can know that there are several hyperparameters in the proposed model. To evaluate the impacts of them, we evaluate two crucial ones among all of them, the number of topics T and the smoothing factor σ. Table 3 shows the influence of the number of topics. From the table, we can observe that the proposed model obtains the best performance when T is set to 30. And performance decreases with more number of topics. We think that data sparsity may be one of the main reasons. With much more topic number, the data sparsity problem will be more serious when

Table 2: Evaluation results of NB, SVM, IBM1, TTM, and our method on the evaluation dataset *RAN-Weibo*.

Methods	Precision	Recall	F_1
NB	0.168	0.154	0.161
SVM	0.312	0,286	0.298
IBM1	0.236	0.214	0.220
TTM	0.356	0.327	0.341
Our w/o BG	0.488	0.448	0.467
Our	**0.522**	**0.479**	**0.500**

Table 3: The influence of the number of topics T of the proposed method.

T	Precision	Recall	F_1
10	0.516	0.473	0.493
30	**0.522**	**0.479**	**0.500**
50	0.508	0.466	0.486
70	0.489	0.449	0.468
100	0.488	0.448	0.467

estimating topic-specific translation probability. Table 4 shows the influence of the translation probability smoothing parameter σ. When σ is set to 0.0, it means that the topical information is omitted. Comparing the results of $\sigma = 0.0$ and other values, we can observe that the topical information can benefit this task. When σ is set to 1.0, it represents the method without smoothing. The results indicate that it is necessary to address the sparsity problem through smoothing.

Figure 2 shows the results of different methods on the dataset *AMB-Weibo*. All the models are trained with same dataset as the above experiments. From the results, we can observe that the F1-scores vary from less than 0.40 up to almost 0.60. The performances' variations of other methods are also huge. We think that training data size and difficulty level are two main reasons. The size of training data of different targets vary greatly in the dataset. However, comparing with other method, the proposed method is the most stable one. Comparing with other methods, the proposed method achieves better performance than other methods for all targets.

4 Related Work

Organization name disambiguation task is fundamental problems in many NLP applications. The task aims to distinguish the real world relevant of a given name with the same surface in context. WePS-3[6] (Amigó et al., 2010) and RepLab 2013[7] (Amigó et al., 2013) evaluation campaigns have also addressed the problem from the perspective of disambiguation organization names in microblogs. Pedersen et al. (2006) proposed an unsupervised method for name discrimination. Yerva et al. (2010) used support vector machines (SVM) classifier with various external resources, such as WordNet, metadata profile, category profile, Google set, and so on. Kozareva and Ravi (2011) proposed to use latent dirichlet allocation to incorporate topical information. Zhang et al. (2012) proposed to use adaptive method for this task. However, most of these methods focused on the text with predefined surface words. The documents which do not contain organization names or person names can not be well processed by these methods.

To bridge the vocabulary gap between content and hashtags, Liu et al. (2012b) proposed to use translation model to handle it. They modeled the tag suggestion task as a translation process from

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

[7] http://www.limosine-project.eu/events/replab2013

Table 4: The influence of the smoothing parameter σ of the propose method.

σ	Precision	Recall	F_1
0.0	0.471	0.432	0.451
0.2	0.490	0.449	0.469
0.4	0.495	0.454	0.474
0.6	0.511	0.468	0.489
0.8	**0.522**	**0.479**	**0.500**
1.0	0.519	0.476	0.496

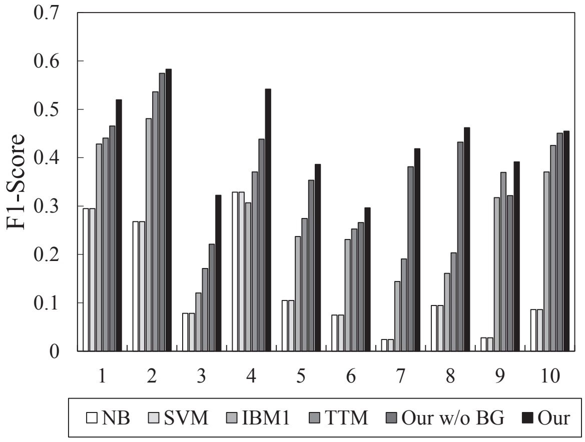

Figure 2: Evaluation results of NB, SVM, IBM1, TTM, and our method on the different companies in the test dataset *AMB-Weibo*.

document content to tags. Ding et al. (2013) extended the translation based method and introduced a topic-specific translation model to process the multiple meanings of words in different topics. Motivated by these methods, we also propose to use topic-specific translation model to handle vocabulary problem. Based on the model, in this work, we incorporate the background knowledge information into the model.

5 Conclusions

To identify target companies of microblogs, in this paper, we propose a novel topical translation model to achieve the task. The main assumption is that the microblog posts and targets describe the same thing with different languages. We convert the target identification problem to a translation process from content of microblogs to targets. We integrate latent topical information into translation model to hand the themes of microblogs in facilitating the translation process. We also incorporate background knowledge (such as product names, series, et al.) into the generation model. Experimental results on a large corpus constructed from a real microblog service and a number of manually labeled golden standards of easily ambiguous entities demonstrate that the proposed method can achieve better performance than other approaches.

6 Acknowledgement

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by 973 Program (2010CB327900), National Natural Science Foundation of China (61003092,61073069), Shanghai Leading Academic Discipline Project (B114) and “Chen Guang” project supported by Shanghai Municipal Education Commission and Shanghai Education Development Foundation(11CG05).

References

Enrique Amigó, Javier Artiles, Julio Gonzalo, Damiano Spina, Bing Liu, and Adolfo Corujo. 2010. Weps3 evaluation campaign: Overview of the on-line reputation management task. In *CLEF (Notebook Papers/LABs/Workshops)*.

Enrique Amigó, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín, Edgar Meij, Maarten Rijke, and Damiano Spina. 2013. Overview of replab 2013: Evaluating online reputation monitoring systems. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, volume 8138 of *Lecture Notes in Computer Science*, pages 333–352. Springer Berlin Heidelberg.

S. Asur and B.A. Huberman. 2010. Predicting the future with social media. In *Proceedings of WI-IAT 2010*.

Hila Becker, Mor Naaman, and Luis Gravano. 2010. Learning similarity metrics for event identification in social media. In *Proceedings of WSDM '10*.

Adam Bermingham and Alan F. Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage? In *Proceedings of CIKM '10*.

Jiang Bian, Yandong Liu, Eugene Agichtein, and Hongyuan Zha. 2008. Finding the right facts in the crowd: factoid question answering over social media. In *Proceedings of WWW '08*.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8.

Zhuoye Ding, Xipeng Qiu, Qi Zhang, and Xuanjing Huang. 2013. Learning topical translation model for microblog hashtag suggestion. In *Proceedings of IJCAI 2013*.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. In *Proceedings of the National Academy of Sciences*, volume 101, pages 5228–5235.

Ido Guy, Naama Zwerdling, Inbal Ronen, David Carmel, and Erel Uziel. 2010. Social media recommendation based on people and tags. In *Proceedings of SIGIR '10*.

Ido Guy, Uri Avraham, David Carmel, Sigalit Ur, Michal Jacovi, and Inbal Ronen. 2013. Mining expertise and interests from social media. In *Proceedings of WWW '13*.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *Proceedings of AAAI'04*.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of ACL-HLT 2011*, Portland, Oregon, USA.

Zornitsa Kozareva and Sujith Ravi. 2011. Unsupervised name ambiguity resolution using a generative model. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, EMNLP '11, pages 105–112, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chia-Jung Lee and W. Bruce Croft. 2013. Building a web test collection using social media. In *Proceedings of SIGIR '13*, SIGIR '13.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of CIKM '09*.

Kang Liu, Liheng Xu, and Jun Zhao. 2012a. Opinion target extraction using word-based translation model. In *Proceedings of EMNLP-CoNLL '12*.

Zhiyuan Liu, Chen Liang, and Maosong Sun. 2012b. Topical word trigger model for keyphrase extraction. In *Proceedings of COLING*.

Kang Liu, Liheng Xu, and Jun Zhao. 2013. Syntactic patterns versus word alignment: Extracting opinion targets from online reviews. In *Proceedings of ACL 2013*, Sofia, Bulgaria.

Takanobu Otsuka, Takuya Yoshimura, and Takayuki Ito. 2012. Evaluation of the reputation network using realistic distance between facebook data. In *Proceedings of WI-IAT '12*, Washington, DC, USA.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Ted Pedersen, Anagha Kulkarni, Roxana Angheluta, Zornitsa Kozareva, and Thamar Solorio. 2006. An unsupervised language independent method of name discrimination using second order co-occurrence features. In *Computational Linguistics and Intelligent Text Processing*, pages 208–222.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of HL-EMNLP 2005*, Vancouver, British Columbia, Canada.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Comput. Linguist.*, 37(1):9–27, March.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860, New York, NY, USA. ACM.

Damiano Spina, Edgar Meij, Maarten de Rijke, Andrei Oghina, Minh Thuong Bui, and Mathias Breuss. 2012. Identifying entity aspects in microblog posts. In *Proceedings of SIGIR '12*.

Damiano Spina, Julio Gonzalo, and Enrique Amigó. 2013. Discovering filter keywords for company name disambiguation in twitter. *Expert Systems with Applications*, 40(12):4986 – 5003.

Yuanbin Wu, Qi Zhang, Xuangjing Huang, and Lide Wu. 2009. Phrase dependency parsing for opinion mining. In *Proceedings of EMNLP 2009*, Singapore.

Surender Reddy Yerva, Zoltn Mikls, and Karl Aberer. 2010. It was easy, when apples and blackberries were only fruits. In Martin Braschler, Donna Harman, and Emanuele Pianta, editors, *CLEF (Notebook Papers/LABs/Workshops)*.

Lei Zhang, Bing Liu, Suk Hwan Lim, and Eamonn O'Brien-Strain. 2010. Extracting and ranking product features in opinion documents. In *Proceedings of COLING '10*.

Shu Zhang, Jianwei Wu, Dequan Zheng, Yao Meng, and Hao Yu. 2012. An adaptive method for organization name disambiguation with feature reinforcing. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 237–245, Bali,Indonesia, November. Faculty of Computer Science, Universitas Indonesia.