Learning Topical Translation Model for Microblog Hashtag Suggestion

Zhuoye Ding, Xipeng Qiu, Qi Zhang, Xuanjing Huang
School of Computer Science, Fudan University
825 Zhangheng Road, Shanghai, P.R. China
{09110240024,xpqiu,qz,xjhuang}@fudan.edu.cn

Abstract

Hashtags can be viewed as an indication to the context of the tweet or as the core idea expressed in the tweet. They provide valuable information for many applications, such as information retrieval, opinion mining, text classification, and so on. However, only a small number of microblogs are manually tagged. To address this problem, in this work, we propose a topical translation model for microblog hashtag suggestion. We assume that the content and hashtags of the tweet are talking about the same themes but written in different languages. Under the assumption, hashtag suggestion is modeled as a translation process from content to hashtags. Moreover, in order to cover the topic of tweets, the proposed model regards the translation probability to be topic-specific. It uses topic-specific word trigger to bridge the vocabulary gap between the words in tweets and hashtags, and discovers the topics of tweets by a topic model designed for microblogs. Experimental results on the dataset crawled from real world microblogging service demonstrate that the proposed method outperforms state-of-the-art methods.

1 Introduction

With the fast growth of Web 2.0, microblogging services have attracted hundreds of millions of web users to publish short instant posts. A hashtag is a string of characters preceded by the symbol (#). In many cases, hashtags can be viewed as an indication to the context of the tweet or as the core idea expressed in the tweet. They can be placed at any point of user wish. If a hashtag is located in the beginning or middle of a tweet, it should be a grammatical part of the tweet. If a hashtag is putted at the end, it needs not to be a grammatical part of the sentence. For example, either "Watching #Avatar.. Great movie!" or "Watching Avatar.. Great movie! #avatar" would be acceptable.

Hashtags have been proven to be useful for many applications, including microblog retrieval [Efron, 2010], query expansion [A.Bandyopadhyay *et al.*, 2011], sentiment analysis [Davidov *et al.*, 2010; Wang *et al.*, 2011]. However, only a small number of tweets are manually tagged. How to automatically generate or recommend hashtags has become an interesting research topic.

In microblogs, posts are usually shorter than traditional documents. Due to the space limit, sometimes hashtags may not appear in the tweet content. To solve the problem of *vocabulary gap*, some approaches based on translation model have been proposed and achieved significant improvement [Liu *et al.*, 2011; Zhou *et al.*, 2011; Bernhard and Gurevych, 2009]. The translation model assumes the content and tags of a resource are describing the same topic but written in different languages. Then it regards tag suggestion as a translation process from document content to tags.

Compared with traditional text collections, suggesting hashtags for microblogs is more challenging for two reasons. First, because of the open access in microblogs, topics tend to be more diverse in microblogs than in formal documents [Zhao *et al.*, 2011b]. Second, due to the 140 character limit, tweets are often much shorter. This makes it extremely hard to determine the topics for lack of sufficient context. Thus, the existing translation model is sometimes vague without the aid of background knowledge. For example, the word "jobs" should be translated into hashtag "Job hunting" in the context of topic *employment*, or "Steve Jobs" under the topic of *Technology*. So an intuitive idea is discovering latent topics of tweets and suggesting hashtags according to the specific topic.

To discover topics for tweets, standard topic model LDA [Blei *et al.*, 2003] was employed in [Ding *et al.*, 2012]. However, the assumptions for long document in LDA may not be satisfied in microblogs. Some existing studies have also proved that LDA cannot deal with the tweet well due to its shortness and sparsity [Zhao *et al.*, 2011b]. Recently, there has been much progress in modeling topics for short texts [Diao *et al.*, 2012; Zhao *et al.*, 2011a; Chen *et al.*, 2011]. Based on these approaches, we introduce a topic model which is more suitable for microblogs in our approach.

In this paper, we propose a topical translation model to recommend hashtags for microblogs. This method regards hashtags and tweet content as *parallel* description of a resource. We integrate latent topical information into translation model to facilitate translation process. On one hand, our model uses topic-specific word trigger to bridge the vocabulary gap between the words in tweets and hashtags. On the other hand, it can determine the topic of tweets by a topic model designed for microblogs. Our model can inherit the advantages of both translation model and topic model.

To demonstrate the effectiveness of our model, we carry our experiments on a large microblogs dataset annotated by users. We find that our model can suggest more appropriate hashtags, compared with some state-of-the-art methods and two degenerate variations of our model.

2 Related Work

Previous work on tag suggestion can be roughly divided into three directions, including collaborative filtering(CF) [Herlocker *et al.*, 2004; Rendle *et al.*, 2009], classification models [Ohkura *et al.*, 2006; Heymann *et al.*, 2008], and generative models[Krestel *et al.*, 2009; Blei and Jordan, 2003].

The collaboration-based methods are typically based on the tagging history of the given resource and user, without considering resource descriptions. FolkRank [Jaschke *et al.*, 2008] and Matrix Factorization [Rendle *et al.*, 2009] were representative collaborative filtering methods for social tag suggestion. Most of these methods suffer from the *cold-start* problem, i.e. they are not able to perform effective suggestions for resources that no one has annotated yet. The content-based approach remedies the *cold-start* problem of the collaboration-based methods by suggesting tags according to content. Therefore, the content-based approach plays an important role in social tag suggestion. The following two directions are content-based models.

Classification models regarded social tag suggestion as a classification problem by considering each tag as a category label. Various classifiers such as Naive Bayes, kNN, SVM and neural networks [Ohkura *et al.*, 2006; Mishne, 2006; Fujimura *et al.*, 2008; Lee and Chun, 2007] have been explored to solve the social tag suggestion problem. To bridge the *vocabulary gap* between content and tags, a new approach based on translation model has been proposed for tag suggestion [Liu *et al.*, 2011]. The translation model regards tag suggestion as a translation process from document content to tags. In order to suggest topic-related tags, a topic-specific translation model was proposed and achieved significant improvement [Ding *et al.*, 2012; Liu *et al.*, 2012]. Our method is mainly based on the study by [Ding *et al.*, 2012]. The main difference is that in their model standard LDA is employed to discover the topics, while our model takes the characteristics of microblogs into consideration and modifies the LDA model. We find the improved topic model can largely boost the performance.

Inspired by the popularity of latent topic models such as Latent Dirichlet allocation(LDA), various generative methods have been proposed to model tags using generative latent topic models. An approach based on Latent Dirichlet allocation was introduced for recommending tags of resources [Krestel *et al.*, 2009]. A LDA-based topic model, Content Relevance Model, was proposed to find the content-related tags for suggestion [Iwata *et al.*, 2009].

Our proposal is complementary to these efforts, because it can integrate the advantages of topic model and translation model. More important, it can capture the characteristics of microblogs well.

Symbol	Description
D	number of annotated tweets
W	number of unique words
T	number of unique hashtags
K	number of topics
N_d	number of words in the dth tweet
M_d	number of hashtags in the dth tweet
z_d	topic of the dth tweet
$w_d = \{w_{dn}\}_{n=1}^{N_d}$	words in the dth tweet
$y_d = \{y_{dn}\}_{n=1}^{N_d}$	topic words or background words
$t_d = \{t_{dm}\}_{m=1}^{M_d}$	hashtags in the dth tweet

Table 1: Notations of our model.

3 Proposed Method

3.1 Preliminaries

We first introduce the notation used in this paper and formally formulate our problem. We assume an annotated corpus consisting of D tweets with a word vocabulary of size W and a hashtag vocabulary of size T. Suppose there are K topics embedded in the corpus. The dth tweet consists of a pair of words and assigned hashtags (w_d, t_d), where $w_d = \{w_{dn}\}_{n=1}^{N_d}$ are N_d words in the tweet that represent the content, and $t_d = \{t_{dm}\}_{m=1}^{M_d}$ are M_d assigned hashtags. Our notation is summarized in Table 1. Given an unlabeled data set, the task of hashtag recommendation is to discover a list of hashtags for each tweet.

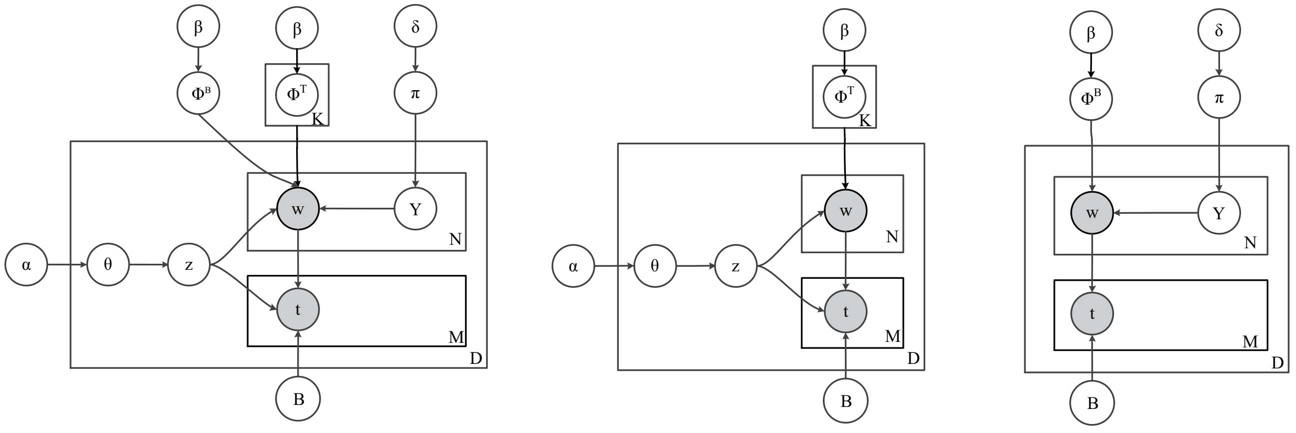
Our method consists of a model learning step and a tag suggestion step. At the model learning step, we learn a topical translation model. At tag suggestion step, we first identify topics for each tweet and then generate candidate hashtags according to the learned model.

3.2 Our Model

As described in Section 1, translation model addresses the problem of vocabulary gap between tweets and hashtags, and can thus suggest hashtags that are uncommon or even not appear in the tweet. Topic model takes the topic of tweets into consideration when suggesting hashtags. In order to aggregate the advantages of both the methods, we propose a topical translation model for hashtag suggestion.

In standard LDA, a document contains a mixture of topics, represented by a topic distribution, and each word has a hidden topic label. While this is a reasonable assumption for long documents, for short microblog posts, a single post is most likely to be about a single topic. We therefore associate a single hidden variable with each post to indicate its topic. Similar ideas of assigning a single topic to a short squence of words has been used before [Diao *et al.*, 2012].

Another observation in microblogs is that posts are noisy and diverse. Besides some "topic words" that are very related to the topic, there are some"background words" that



(a)

(b)

(c)

Figure 1: (a) Our topical translation model for hashtag suggestion(TTM). (b) A variation of our model where we consider all the words are topic words(TTM_1). (c) A variation of our model where we neglect the topical information(TTM_2).

are commonly used in tweets on different topics. We assume that words are generated either from a topic-specific distribution or from a corpus-wide background distribution. We use $y_d = \{y_{dn}\}_{n=1}^{N_d}$ to indicate a word to be a topic word or a background word. Moveover, we assume a background word distribution ϕ_B that captures background words. All posts are assumed to be generated from some mixture of these $K + 1$ underlying topics.

The proposed topical translation model is based on the following assumptions. When a user wants to write a tweet, he first generates the content, and then generates the hashtags. When starting the content, he first chooses a topic based on the topic distribution. Then he chooses a bag of words one by one from the word distribution for the topic or from the background word distribution that captures white noise. During the generative process for hashtags, hashtags are chosen according to the topic and topic words in the content.

Formally, we use π to denote the probability of choosing to a topic word or a background word. Let θ denote the topic distribution and ϕ_k denote the word distribution for topic k. ϕ_B denotes the word distribution for background words. And then each hashtag t_{dm} is annotated according to topic-specific translation possibility $P(t_{dm}|w_d, z_d, \mathbf{B})$, where $\mathbf{B}$ presents the topic-specific word alignment table between a word and a hashtag, where $B_{i,j,k} = P(t = t_j|w = w_i, z = k)$ is the word alignment probability between the word w_i and the hashtag t_j for topic k. In summary, the generation process of annotated tweets is described in Algorithm 1. Figure 1(a) shows a graphical model representation of the proposed model.

There are two degenerate variations of our model that we also consider in our experiments. The first one is depicted in Figure 1(b). In this model, we consider all the words in a tweet are topic words and generated from a topic-specific distribution. So the hidden variable y is neglected. The second one is depicted in Figure 1(c). In this model, we neglect the topic information. We refer to our complete topic translation model as TTM Model, the model in Figure 1(b) as TTM_1 Model and the model in Figure 1(c) as TTM_2 Model. In Section 3.3, we describe the learning details of our model. Due to the space limit, we leave out the learning details of two degenerate variations of our model which are very similar to our model.

Algorithm 1 The Generation Process of Annotated Tweets

```
Draw $\pi \sim$ Beta$(\delta)$
Draw background word probability $\phi_B \sim$ Dir$(\beta)$
**for all** topic $k = 1, ..., K$ **do**
    Draw topic word probability $\phi_k \sim$ Dir$(\beta)$
**end for**
**for all** tweet $d = 1, ..., D$ **do**
    Draw topic probability $\theta_d \sim$ Dir$(\alpha)$
    Draw topic $z_d \sim$ Multi$(\theta_d)$
    **for all** word $n = 1, ..., N_d$ **do**
        Draw $y_{dn} \sim$ Bernoulli$(\pi)$
        **if** $y_{dn} = 1$ **then**
            Draw word $w_{dn} \sim$ Multi$(\phi_{z_d})$
        **end if**
        **if** $y_{dn} = 0$ **then**
            Draw word $w_{dn} \sim$ Multi$(\phi_B)$
        **end if**
    **end for**
    **for all** hashtag $m = 1, ..., M_d$ **do**
        Draw $t_{dm} \sim P(t_{dm}|w_d, z_d, \mathbf{B})$
    **end for**
**end for**
```

3.3 Learning

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.

The sampling probability of being a topic/background word for ith word in the dth tweet is sampled from:

$$P(y_{di} = p|\mathbf{W}, \mathbf{T}, \mathbf{Z}, \mathbf{Y}_{-di}) \propto \frac{N_{-i,p} + \delta}{N_{-i,.} + 2\delta} \cdot \frac{N_{-i,l}^{w_{di}} + \beta}{N_{-i,l}^{(\cdot)} + \beta W}$$

where $l = B$ when $p = 0$ and $l = z_d$ when $p = 1$. $N_{-i,p}$ is a count of words that are assigned to background words and any topic respectively. $N_{-i,B}^{w_{di}}$ is the number of w_{di} that assigned to background words. $N_{-i,z_d}^{w_{di}}$ is the number of w_{di} that are assigned to topic z_d. All counters are calculated with the current word w_{di} excluded.

The sampling probability of a latent topic for the dth tweet is sampled from:

$$P(z_d = k|\mathbf{W}, \mathbf{T}, \mathbf{Z}_{-d}, \mathbf{Y}) \propto \frac{N_{-d,k} + \alpha}{N_{-d,.} + \alpha K} \cdot \prod_{i=0}^{N_d} \frac{N_{-d,k}^{w_{di}} + \beta}{N_{-d,k}^{(\cdot)} + \beta W} \cdot \prod_{j=0}^{M_d} \sum_{i=0}^{N_d} \frac{M_{-d,k}^{w_{di} t_{dj}} + \beta}{M_{-d,k}^{(w_{di} \cdot)} + \beta T}$$

Where $N_{-d,k}$ is a count of tweets that are assigned topic k in the corpus. $N_{-d,k}^{w_{di}}$ is a count of topic words w_{di} that are assigned to topic k in the corpus, here topic words refer to words whose latent variable y equals 1. $M_{-d,k}^{w_{di} t_{dj}}$ is the number of occurrences that word w_{di} is translated to hashtag t_{dj} given topic k. All counters with $-d$ are calculated with the current tweet w_d excluded.

After all the hidden variables become stable, we can estimate topic-specific word alignment table B by: $B_{t,w,z} = \frac{N_{z,w}^{t}}{N_{z,w}^{(\cdot)}}$. where $N_{z,w}^{t}$ is a count of the hashtag t that co-occurs with the word w under topic z in tweet-hashtag pairs.

The possibility table $B_{t,w,z}$ have a potential size of $W \cdot T \cdot K$, assuming the vocabulary sizes for words, hashtags and topics are W, T and K. The data sparsity poses a more serious problem in estimating $B_{t,w,z}$ than the topic-free word alignment case. To reduce the data sparsity problem, we introduce the remedy in our model. We can employ a linear interpolation with topic-free word alignment probability to avoid data sparsity problem:

$$B_{t,w,z}^{*} = \lambda B_{t,w,z} + (1 - \lambda) P(t|w)$$

where $P(t|w)$ is topic-free word alignment probability between the word w and the hashtag t. Here we explore IBM model-1 [Brown *et al.*, 1993], which is a widely used word alignment model, to obtain $P(t|w)$. λ is trade-off of two probabilities ranging from 0.0 to 1.0. When $\lambda = 0.0$ $B_{t,w,z}^{*}$ will be reduce to topic-free word alignment probability; and when $\lambda = 1.0$, there will be no smoothing in $B_{t,w,z}^{*}$.

3.4 Hashtag Suggestion

Suppose given an unlabeled dataset, we first discover the topic and determine topic/background words for each tweet. The collapsed Gibbs sampling is also applied for inference. The process is the same as described in Section 3.3.

After the hidden variables of topic/background words and the topic of each tweet become stable, we can estimate the distribution of topics for the dth tweet in unlabeled data by: $\eta_{dk}^{*} = \frac{p(k)p(w_{d1}|k)...p(w_{dN_d}|k)}{Z}$, where $p(w_{di}|k) = \frac{N_k^{w_{di}}+\beta}{N_k^{(\cdot)}+\beta W}$ and $N_k^{w_{di}}$ is a count of words w_{di} that are assigned to topic k in the corpus. And $p(k) = \frac{N_k+\alpha}{N_{.}+\alpha K}$ is regarded as a prior for topic distribution, where N_k is a count of tweets that are assigned to topic k. Z is the normalized factor.

With topic distribution η^* and topic-specific word alignment table B^*, we can rank hashtags for the dth tweet in unlabeled data by computing the scores:

$$P(t_{dm}|w_d, \eta_d^*, \mathbf{B}^*) \propto \sum_{z_d=1}^{K} \sum_{n=1}^{N_d} P(t_{dm}|z_d, w_{dn}, \mathbf{B}^*) \cdot P(z_d|\eta_d^*) \cdot P(w_{dn}|w_d)$$

where $p(w_{dn}|w_d)$ is the weight of the word w_{dn} in the tweet content w_d, which can be estimated by the IDF score of the word. According to the ranking scores, we can suggest the top-ranked hashtags for each tweet to users.

4 Experiments

4.1 Dataset and Settings

In our experiments, we use a Microblog dataset collected from Sina-Weibo[1] for evaluation. Sina-Weibo is a Twitter-like microblogging system in China provided by Sina, one of the largest Chinese Internet content providers. We collect 110,000 tweets that contain hashtags annotated by users. Some detailed statistical information is shown in Table 2. Among them, we randomly select 10,000 tweets as the test set, and use the rest of the dataset as training set. For evaluation, we use hashtags annotated by users as the golden set.

#tweet	W	T	N_w	N_t
110,000	106,323	37,224	20.45	1.20

Table 2: Statistical information of dataset. W, T, $\bar{N_w}$ and $\bar{N_t}$ are the vocabulary of words, the vocabulary of hashtags, the average number of words in each tweet and the average number of hashtags in each tweet respectively.

We use Precision(P), Recall(R), and F-value(F) to evaluate the performance of hashtag recommendation methods. Precision means the percentage of "tags truly assigned" among "tags assigned by system". Recall means that "tags truly assigned" among "tags manually assigned". F-value is the average of Precision and Recall. We ran our model with 500 iterations of Gibbs sampling. After trying a few different numbers of topics, we empirically set the number of topics to 30. We use $\alpha = 50.0/K$ and $\beta = 0.1$ as [Griffiths and Steyvers, 2004] suggested. Parameter δ is set to 0.01. We set smoothing parameter λ to 0.8 which gives the best performance. The influence of parameters to our model can be found in Section 4.3.

4.2 Evaluation Results

In this section, we compare our topical translation model (Figure 1(a)) with some baseline models and two degeneration models. We consider the following methods.

[1] http://weibo.com/

- **Naive Bayes(NB)**: We formulated hashtag suggestion as a classification task and applied a classification method(NB) to model the posterior probability of each hashtag given a tweet.
- **LDA**: LDA model is applied to recommend hashtags in [Krestel *et al.*, 2009].
- **IBM1**: Translation model(IBM model-1) is applied to obtain the alignment probability between the word and the tag [Liu *et al.*, 2011].
- **TopicWA**: TopicWA is a topical word alignment model, in which standard LDA is employed to discover the latent topic [Ding *et al.*, 2012].
- **TTM_1**: As depicted in Figure 1(b), TTM_1 is a degenerate variation of our model, in which we consider all the words in tweets are topic-related.
- **TTM_2**: As depicted in Figure 1(c), TTM_2 is a degenerate variation of our model, in which we neglect the influence of topic information.

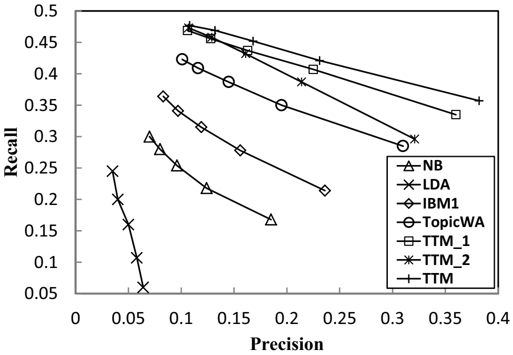


Figure 2: Performance comparison between NB, LDA, IBM1, topicWA, TTM_1, TTM_2, and TTM

In Figure 2, we show the Precision-Recall curves of NB, LDA, IBM1, topicWA, TTM_1, TTM_2 and TTM on the data set. Each point of a Precision-Recall curve represents different numbers of suggested hashtags from M = 1(bottom right, with higher Precision and lower Recall) to M = 5(upper left, with higher Recall but lower Precision) respectively. The closer the curve to the upper right, the better the overall performance of the method. From the Figure, we have the following observations: (1) Our proposed models (TTM_1, TTM_2, TTM) outperform all the baseline methods. This indicates the robustness and effectiveness of our approaches. (2) TopicWA outperforms all the other baselines, because it can combine the advantages of both translation model and topic model. However, TopicWA underperforms TTM, because TopicWA applies standard LDA to discover the topics. And LDA cannot perform well in modeling the topics for microblogs. While TTM inherits all the advantages of TopicWA, moreover, it applies an improved topic model designed for microblogs. (3) LDA performs so poor, because it ranks the candidate hashtags by the topic-hashtag distribution. So it can only suggest general tags. It is congruence with the previous conclusion given by [Ding *et al.*, 2012]. (4) TTM can also outperform the two degenerate variations of the model, which proves the consideration of topic/background words and topic information are both helpful. TTM_1 performs better than TTM_2. This indicates that topic information is more important for suggesting hashtags.

To further demonstrate the performance of TTM and other baseline methods, in Table 3, we show the Precision, Recall and F-measure of NB, LDA, IBM1, TopicWA, and our proposed models suggesting top-1 hashtag, because the number is near the average number of hashtags in dataset. We find that the F-measure of TTM comes to 0.364, outperforming all the baselines more than 7%.

Method	Precision	Recall	F-measure
NB	0.217	0.197	0.203
LDA	0.064	0.060	0.062
IBM1	0.236	0.214	0.220
TopicWA	0.310	0.285	0.292
TTM_1	0.360	0.335	0.343
TTM_2	0.321	0.296	0.303
TTM	**0.382**	**0.357**	**0.364**

Table 3: Comparison results of NB, LDA-based, IBM1, TopicWA, TTM_1, TTM_2, and TTM.

K	Precision	Recall	F-measure
K=10	0.353	0.328	0.335
K**=30**	**0.382**	**0.357**	**0.364**
K=50	0.365	0.341	0.348
K=70	0.356	0.331	0.339
K=100	0.340	0.316	0.323

Table 4: The influence of topic number K of TTM for hashtag suggestion.

4.3 Parameter Influence

There are two crucial parameters in our model, the number of topics K and the smoothing factor λ. In this section, we demonstrate the performance of our model for hashtag suggestion when parameters change.

In Table 4, we find that our model obtains the best performance with 30 topics. And performance decreases with a small topic size. Because a small number of topics typically leads to fairly general topics. Such general topics have a higher chance to suggest topic-unrelated hashtags. On the contract, with much more topic number, the data sparsity problem will be more serious when estimating topic-specific translation probability.

As shown in Figure 3, when the smoothing factor is set to $\lambda = 0.8$, our model achieves the best performance. When either $\lambda = 0.0$ and $\lambda = 1.0$, the performance is poorer compared to smoothed model. This reveals that it is necessary to address the sparsity problem of our model by smoothing.

4.4 Sample Results and Discussion

Besides quantitative evaluation, we investigate the topics learned by our model. As shown in Table 5, we select two

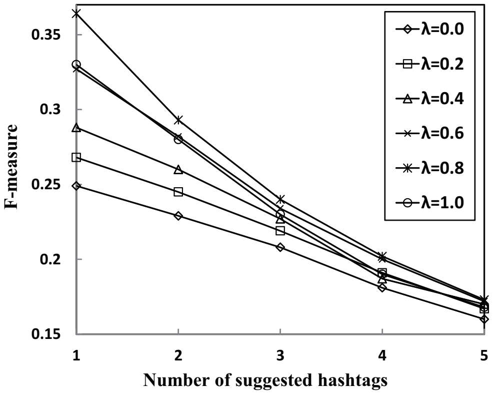


Figure 3: F-measure of TTM when smoothing parameter λ ranges from 0.0 to 1.0.

Topic	Top words	$P(T\|w=$"Apple"$,z)$
Topic-9	Food, fruit, apple, milk banana	Fruit, Lose weight, Apple, Diet, Health knowledge
Topic-19	advertisement, internet, user, mobile, platform	Apple, iphone, iphone skills, Android, Technology

Table 5: Examples of topics learned by our model and the translation probabilities with respect to the topics

topics, i.e., Topic-9 and Topic-19 for study. In the second column of the Table, we list the top-5 words given by the two topics separately (ranked by $p(w|z)$) . From the top words, we can conclude that Topic-9 is about "Food and health" and Topic-19 is about "Technology" .

To further analyse the translation probabilities between words and hashtags under different topics. We pick a word "Apple" for example. In the last column of the Table, we show the top-5 hashtags triggered by the word "Apple" with respect to the two topics according to the probability $p(t|w,z)$. We can see that they are discriminative with each other. In the context of the topic "Food and health", the word "Apple" generally refers to "fruits" and thus correlates to some hashtags about foods and health; while in the context of the topic "Technology", the word "Apple" always correlates with hashtags "iphone" and "Android".

After investigating the topics, we look into hashtags suggested by IBM1, TopicWA, TTM_1, TTM_2 and TTM for a tweet. Here we select a tweet "After the earthquake in Aobaku Sendai Japan on March 11, 2011, people escaped in the green belt to make vehicles driven on the road." for example. In Table 6, we show the top-5 hashtags, in which we use (-) to highlight the inappropriate hashtags.

From Table 6, we observe that IBM1 method suggests some topic-unrelated hashtags, because it relies solely on word-tag co-occurrence statistics without considering the topic. For instance, "house" is triggered by the words "road" and "green belt". Despite considering the topic, TopicWA still suggests some inappropriate hashtags, because standard LDA cannot do well in modeling topics for microblogs. On the contrary, our proposed models (TTM, TTM_1, TTM_2) achieve improvement. Because they take the characteristics of microblogs into consideration when modeling topics. And our complete model TTM obtains the best performance, and all the hashtags suggested by TTM are topic-related and appropriate.

Method	Suggested hashtags
IBM1:	Japan earthquake, House(-), Japan earthquake map, Shanghai culture(-), Share picture(-),
TopicWA:	Japan earthquake, Help, Japan earthquake map, House(-), Fukushima nuclear crisis(-)
TTM_1:	Japan earthquake, Japan earthquake tsunamis, Help, Paintings(-), 9.0 earthquake Japan
TTM_2:	Japan earthquake, Japan earthquake tsunamis, 9.0 earthquake Japan, Help, Situation in Libya(-)
TTM:	Japan earthquake, Japan earthquake tsunamis, Japan earthquake map, 9.0 earthquake Japan, help

Table 6: Top-5 hashtags suggested by IBM1, TopicWA, TTM_1, TTM_2 and TTM

5 Conclusions and Future Work

To suggest hashtags for microblogs, in this paper, we proposed a topical translation model, which combines the advantages of both topic model and translation model. On one hand, our model uses topic-specific word trigger to bridge the vocabulary gap between the words and hashtags. On the other hand, it can discover the topics of tweets by a topic model designed for microblogs. Experimental results on the dataset crawled from real world microblogging service demonstrate that the proposed method outperforms state-of-the-art methods.

We design the following research plans: (1) TTM does not take the information of users into consideration. We plan to incorporate user information into our topic model. Furthermore, social network information can also been applied in our model. (2) We demonstrate the utility of our approach in microblogs sites. In the future, we also plan to focus on applying the techniques to other social media.

Acknowledgments

The author wishes 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 Science and Technology Development Funds(12511504502), Shanghai Leading Academic Discipline Project (B114), and "Chen Guang" project supported by Shanghai Municipal Education Commission and Shanghai Education Development Foundation (11CG05).

References

[A.Bandyopadhyay *et al.*, 2011] A.Bandyopadhyay, M. Mitra, and P. Majumder. Query expansion for microblog retrieval. In *Proceedings of The Twentieth Text REtrieval Conference*, TREC 2011, 2011.

[Bernhard and Gurevych, 2009] Delphine Bernhard and Iryna Gurevych. Combining lexical semantic resources with question & answer archives for translation-based answer finding. In *Proceeding of ACL*, pages 728–736, 2009.

[Blei and Jordan, 2003] D.M. Blei and M.I. Jordan. Modeling annotated data. In *Proceedings of SIGIR*, pages 127–134, 2003.

[Blei *et al.*, 2003] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[Brown *et al.*, 1993] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The machematics of statistical machine translation: parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

[Chen *et al.*, 2011] Mengen Chen, Xiaoming Jin, and Dou Shen. Short text classfication improved by learning multigranularity topics. In *proceedings of IJCAI*, 2011.

[Davidov *et al.*, 2010] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 241–249, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Diao *et al.*, 2012] Qiming Diao, Jing Jiang, Feida Zhu, and Ee-Peng Lim. Finding bursty topics from microblogs. In *Proceedings of ACL*, 2012.

[Ding *et al.*, 2012] Zhuoye Ding, Qi Zhang, and Xuanjing Huang. Automatic hashtag recommendation for microblogs using topic-specific translation model. In *Proceeding of COLING*, 2012.

[Efron, 2010] Miles Efron. Hashtag retrieval in a microblogging environment. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 787–788, New York, NY, USA, 2010. ACM.

[Fujimura *et al.*, 2008] S. Fujimura, KO Fujimura, and H. Okuda. Blogosonomy: Autotagging any text using bloggers' knowledge. In *Proceedings of WI*, 2008.

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

[Herlocker *et al.*, 2004] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, January 2004.

[Heymann *et al.*, 2008] Paul Heymann, Daniel Ramage, and Hector Garcia-Molina. Social tag prediction. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 531–538, New York, NY, USA, 2008. ACM.

[Iwata *et al.*, 2009] T. Iwata, T. Yamada, and N. Ueda. Modeling social annotation data with content relevance using a topic model. In *Proceedings of NIPS*, pages 835–843, 2009.

[Jaschke *et al.*, 2008] R. Jaschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in social bookmarking systems. *AI Communications*, 21(4):231–247, 2008.

[Krestel *et al.*, 2009] Ralf Krestel, Peter Fankhauser, and Wolfgang Nejdl. Latent dirichlet allocation for tag recommendation. In *RecSys*, 2009.

[Lee and Chun, 2007] S.O.K. Lee and A.H.W. Chun. Automatic tag recommendation for the web 2.0 blogosphere using collaborative tagging and hybrid ann semantic structures. In *Proceedings of WSEAS*, pages 88–93, 2007.

[Liu *et al.*, 2011] Zhiyuan Liu, Xinxiong Chen, and Maosong Sun. A simple word trigger method for social tag suggestion. In *Proceedings of EMNLP*, 2011.

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

[Mishne, 2006] G. Mishne. Autotag: a collaborative approach to automated tag assignment for weblog posts. In *Proceedings of WWW*, pages 953–954, 2006.

[Ohkura *et al.*, 2006] Tsutomu Ohkura, Yoji Kiyota, and Hiroshi Nakagawa. Browsing system for weblog articles based on automated folksonomy. *Workshop on the Weblogging Ecosystem Aggregation Analysis and Dynamics at WWW*, 2006.

[Rendle *et al.*, 2009] Steffen Rendle, Leandro Balby Marinho, Alexandros Nanopoulos, and Lars Schmidt-Thieme. Learning optimal ranking with tensor factorization for tag recommendation. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 727–736, New York, NY, USA, 2009. ACM.

[Wang *et al.*, 2011] Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 1031–1040, New York, NY, USA, 2011. ACM.

[Zhao *et al.*, 2011a] Wayne Xin Zhao, Jing Jiang, and Jing He. Topical keyphrase extraction from twitter. In *Proceedings of ACL*, 2011.

[Zhao *et al.*, 2011b] Wayne Xin Zhao, Jing Jiang, and Jianshu Weng. Comparing twitter and traditional media using topic models. In *Proceedings of ECIR*, 2011.

[Zhou *et al.*, 2011] Guangyou Zhou, Li Cai, Jun Zhao, and Kang Liu. Phrase-based translation model for question retrieval in community question answer archives. In *Proceeding of ACL*, pages 653–662, 2011.